

NYO-1480-63

SEP 21 1966

Courant Institute of
Mathematical Sciences

NEW YORK UNIVERSITY
COURANT INSTITUTE - LIBRARY
251 Mercer St. New York, N.Y. 10012

AEC Computing and Applied Mathematics Center

Numerical Analysis of
Boundary Layer Problems

William D. Murphy

AEC Research and Development Report

TID-4500
MATHEMATICS
August 1966

New York University



NYO-1480-63

C.I.

NEW YORK UNIVERSITY
COURANT INSTITUTE - LIBRARY
251 Mercer St. New York, N.Y. 10012

UNCLASSIFIED

AEC Computing and Applied Mathematics Center
Courant Institute of Mathematical Sciences
New York University

TID-4500

NYO-1480-63
Mathematics

NUMERICAL ANALYSIS OF BOUNDARY LAYER PROBLEMS

by

William D. Murphy

Contract No. AT(30-1)-1480

UNCLASSIFIED

NEW YORK UNIVERSITY
COURANT INSTITUTE - LIBRARY

Abstract

We categorize some of the finite difference methods that can be used to treat the initial value problem for the boundary layer differential equation $\mu y' = f(y, x)$. These methods take the form

$$\sum_{i=0}^k \alpha_i Y_{n+i} = h^{1-\gamma} \sum_{i=0}^k \beta_i f(Y_{n+i}, x_{n+i}),$$

where $\mu = h^\gamma$, $0 < \gamma < 1$, and h is the meshsize. We define a new kind of stability called μ -stability and prove that under certain conditions μ -stability implies convergence of the difference method. We investigate μ -stability and the optimal methods which it allows (i.e. methods of maximum accuracy). These results are generalized to the vector case.

We also consider the related finite difference methods for the initial value problem

$$\begin{aligned} \mu y' &= f(y, z, x) \\ z' &= g(y, z, x). \end{aligned}$$

To illustrate the theoretical results, a number of problems were solved by these methods with the use of New York University's CDC6600 and IBM709⁴ computers.

ACKNOWLEDGEMENT

I would like to express my deep sense of gratitude to Professor Eugene Isaacson for spending many hours discussing and reading this report and for making a number of valuable suggestions and comments.

WDM

TABLE OF CONTENTS

Chapter 1	Numerical Analysis and Boundary Layer Problems	1
Chapter 2	Linear Equations and μ -Stability	15
Chapter 3	Optimal Methods	42
Chapter 4	Boundary Layer Equations of the First Order in One Unknown	52
Chapter 5	Extension to Higher Dimensions	88
Chapter 6	Boundary Layer Equations of Higher Order	115
Chapter 7	Applications and Final Statments	129
Bibliography		139

I. NUMERICAL ANALYSIS AND BOUNDARY LAYER PROBLEMS

Numerical methods for solving the initial value problem on the interval $[0,1]$ for the ordinary differential equation (ODE) problem

$$(1.1) \quad y' = f(y,x) \quad \text{and} \quad y(0) = 1$$

are well understood. Among the most useful methods are the finite difference schemes. For these, we divide the interval $[0,1]$ into mesh points at a distance h apart: $x_0 = 0$, $x_1 = h$, $x_2 = 2h, \dots$, $x_n = nh$. An important class of finite difference methods can be written as

$$(1.2) \quad \sum_{i=0}^k \alpha_i Y_{n+i} = h \sum_{i=0}^k \beta_i f(Y_{n+i}, x_{n+i})$$

where k is a fixed integer and α_v and β_v ($v = 0, 1, \dots, k$) denote real constants which do not depend upon n . The first proof that the solution of equation (1.2) converges to the solution of equation (1.1) if and only if certain algebraic conditions are satisfied was given by Dahlquist [1]. Additional literature may be found in Henrici [2] and Hull and Luxemburg [3]. We will describe this theory briefly in order to prepare the reader for the

more complicated numerical analysis of boundary layer equations.

We associate with the difference equation (1.2) two polynomials

$$\rho(\xi) = \alpha_k \xi^k + \alpha_{k-1} \xi^{k-1} + \dots + \alpha_0$$

$$\sigma(\xi) = \beta_k \xi^k + \beta_{k-1} \xi^{k-1} + \dots + \beta_0.$$

We assume for convenience that $\rho(\xi)$ and $\sigma(\xi)$ have no common factors. In order that the difference equation be a good approximation to the ODE $y' = f(y, x)$ we must prescribe certain consistency conditions on the coefficients α_μ and β_μ .

DEFINITION. A difference equation of the form (1.2) is said to be consistent if

$$a) \quad \rho(1) = 0$$

and

$$b) \quad \rho'(1) = \sigma(1).$$

We must also ensure that the difference equation be insensitive to small changes in the local errors. This will be guaranteed, as shown in [1], by the stability condition.

DEFINITION. A difference equation of the form (1.2) is said to be stable if the roots of $\rho(\zeta) = 0$ lie in or on the unit circle in the complex ζ -plane, and are simple if they lie on the circle.

Since (1.2) is a k^{th} order difference scheme, we must supply k initial values $Y_0, Y_1, \dots, Y_{k-1}$ in order to solve the equation. We may let $Y_0 = y(0) = 1$ and calculate by Taylor series, Runge-Kutta, or some other method the values of $Y_1, Y_2, \dots, Y_{k-1}$. We merely prescribe that the method chosen to calculate the initial data allows us to write

$$(1.3) \quad |Y_v - y(x_v)| = \mathcal{O}(h) \quad \text{as } h \rightarrow 0 \quad \text{for } v = 0, 1, \dots, k-1.$$

The more general difference equation

$$(1.4) \quad \sum_{i=0}^k \alpha_i Y_{n+i} = h \sum_{i=0}^k \beta_i f(Y_{n+i}, x_{n+i}) + R_n$$

may also be considered.

Here, R_n is the round-off error and arises from the inaccuracy of the arithmetic process used to solve the difference equation. To guarantee convergence when using equation (1.4) we add the condition that

$$R_n = \mathcal{O}(h^2) \quad \text{as } h \rightarrow 0.$$

With the difference equations (1.2) or (1.4) we associate the difference operator

$$(1.5) \quad L[y(x);h] = \sum_{i=0}^k [\alpha_i y(x+ih) - h\beta_i y'(x+ih)].$$

We will assume the function $y(x)$ has continuous derivatives of sufficiently high order and expand $y(x+ih)$ and $y'(x+ih)$ in Taylor series:

$$y(x+ih) = y(x) + ihy'(x) + \frac{1}{2} i^2 h^2 y''(x) + \dots$$

$$hy'(x+ih) = hy'(x) + ih^2 y''(x) + \dots$$

Substituting these expressions into equation (1.5) gives

$$\begin{aligned} L[y(x);h] &= C_0 y(x) + C_1 hy'(x) + C_2 h^2 y''(x) \\ &\quad + \dots + C_q h^q y^{(q)}(x) + \dots \end{aligned}$$

where

$$\begin{aligned} C_q &= \frac{1}{q!} (\alpha_1 + 2^q \alpha_2 + \dots + k^q \alpha_k) \\ &\quad - \frac{1}{(q-1)!} (\beta_1 + 2^{q-1} \beta_2 + \dots + k^{q-1} \beta_k). \end{aligned}$$

We note that by the consistency condition $C_0 = \rho(1) = 0$ and $C_1 = \rho'(1) - \sigma(1) = 0$. A difference operator of the form (1.5) is said to be of order p if $C_0 = C_1 = \dots = C_p = 0$, but $C_{p+1} \neq 0$. In this case we may write

$$L[y(x);h] = \mathcal{O}(h^{p+1}) \quad \text{as } h \rightarrow 0$$

if the derivatives are also bounded. The associated difference equation satisfied by the solution of (1.1) is

$$\sum_{i=0}^k \alpha_i y_{n+i} = h \sum_{i=0}^k \beta_i f_{n+i} + T_n$$

where T_n is the truncation error ($T_n = \mathcal{O}(h^{p+1})$ as $h \rightarrow 0$), $y_{n+i} = y(x_{n+i})$, and $f_{n+i} = f(y(x_{n+i}), x_{n+i})$. We are now in a position to define convergence.

DEFINITION. A difference equation of the form (1.2) or (1.4) is said to be convergent if

$$\lim_{\substack{h \rightarrow 0 \\ x_n = nh}} Y_n = y(x_n) \quad \text{for all } x_n \in [0,1].$$

The main theorem, first proved by Dahlquist [1], is

THEOREM 1.1. A necessary and sufficient condition for convergence of the consistent finite difference equation given by (1.4) is stability.

Proof: See [1], [2], or [3].

Remark 1. We have assumed that at least $y''(x)$ is continuous and bounded. While less stringent conditions may be used for the equation (1.1), this will not be the case when we study the boundary layer equation.

Remark 2. Although we only considered the scalar equation (1.1), the same theory is applicable to the vector equation $y' = f(y, x)$ where y' and f are s -dimensional vectors.

Boundary layer phenomena may be associated with the equation

$$(1.6) \quad \mu y' = f(y, x) \quad \text{and} \quad y(0) = 1 \quad \text{for} \quad x \in [0, 1],$$

where μ is a small parameter. A number of questions can be posed for such a system. For example,

- a) What is the nature of such a solution for small μ ?
- b) As $\mu \rightarrow 0$ does the solution $y(x, \mu)$ to equation (1.6) approach a solution to the reduced equation

$$(1.7) \quad f(y, x) = 0,$$

and if so, which one?

- c) Can a uniform asymptotic expansion be found for the solution of (1.6)?

These questions have all been answered in recent years not only for the equation (1.6) but also for the vector system

$$\begin{aligned} \mu \frac{dy}{dx} &= f(y, z, x) & y(0) &= y^0 \\ (1.7) \quad \frac{dz}{dx} &= g(y, z, x) & z(0) &= z^0 \end{aligned}$$

$$0 \leq x \leq 1$$

where y is an M -dimensional vector and z is an m -dimensional vector. We will briefly describe the treatment found in Vasil'eva [4].

We first introduce some definitions. Let $y = \phi(z, x)$ be one of the solutions of the system $f(y, z, x) = 0$. By the degenerate system of equations corresponding to the root $y = \phi(z, x)$ we will mean

$$\begin{aligned} y &= \phi(z, x) & 0 \leq x \leq 1 \\ (1.8) \quad \frac{dz}{dx} &= g(y, z, x) & z(0) &= z^0. \end{aligned}$$

Let the solution of this system be denoted by $\bar{z}(x)$ and $\bar{y}(x) = \phi(\bar{z}(x), x)$.

DEFINITION. The root $y = \phi(z, x)$ is isolated on a bounded closed set $D = \{(z, x)\}$ if there exists an $\varepsilon > 0$

such that the system $f(y, z, x) = 0$ has no solution other than $\phi(z, x)$ for $|y - \phi(z, x)| < \epsilon$.

Example 1. Suppose $f(y, z, x) = -y^2 + 1$. The roots are $y = \pm 1$ and are obviously isolated.

Example 2. Suppose $f(y, z, x) = -y^2 + x^2$. The roots are $y = \pm x$. These roots are not isolated at $x = 0$ if we are considering the interval $[-1, 1]$ but are isolated on the interval $[\frac{1}{2}, 1]$.

The equations

$$(1.9) \quad \frac{dy}{d\tau} = f(y, z^*, x^*)$$

(z^* and x^* are regarded as parameters) will be called the adjoined system of equations.

DEFINITION. The isolated root $y = \phi(z, x)$ will be called positively stable in D if the real parts of the roots of the characteristic equation

$$\text{Det}\left(\frac{\partial f}{\partial y}(\phi(z, x), z, x) - \lambda I\right) = 0$$

($\frac{\partial f}{\partial y}$ is a matrix) are negative in D .

DEFINITION. The domain of influence of an isolated positively stable root $y = \phi(z, x)$ is the set of points (y^*, z^*, x^*) such that the solution to the adjoined system

(1.9) satisfying the initial conditions $y|_{\tau=0} = y^*$ tends to the value $\phi(z^*, x^*)$, as $\tau \rightarrow \infty$.

We can now state the main theorem about boundary layer equations:

THEOREM 1.2. If some root $y = \phi(z, x)$ of the system $f(y, z, x) = 0$ is an isolated positively stable root in some bounded closed domain D , if the initial point $(y^0, z^0, 0)$ belongs to the domain of influence of this root, and if the solution $z = \bar{z}(x)$ of the degenerate system (1.8) belongs to D for $0 \leq x \leq X$, then the solution $\begin{pmatrix} y(x, \mu) \\ z(x, \mu) \end{pmatrix}$ of the original system (1.7) tends to the solution $\begin{pmatrix} \phi(\bar{z}(x), x) \\ \bar{z}(x) \end{pmatrix}$ of the degenerate system, as $\mu \rightarrow 0$, the passage to the limit

$$\lim_{\mu \rightarrow 0} y(x, \mu) = \bar{y}(x) \equiv \phi(\bar{z}(x), x)$$

holding for $0 < x \leq 1 < X$, and the passage to the limit

$$\lim_{\mu \rightarrow 0} z(x, \mu) = \bar{z}(x)$$

for $0 \leq x \leq 1 < X$.

Proof: See Vasil'eva [4].

We illustrate the theorem by considering a few examples:

Example 3.

$$\mu y' = -y \quad y(0) = 1.$$

The only root is $y = 0$ and therefore it is isolated. The solution is $y = e^{-x/\mu}$. The graph of this solution is given in Figure 1. The dotted curves show

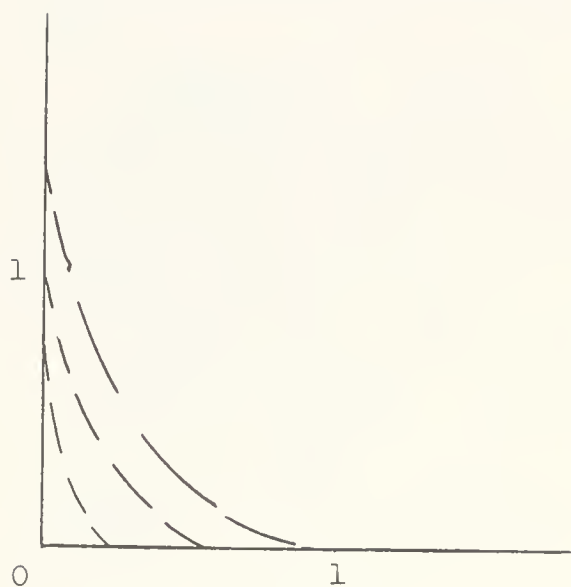


Fig. 1

the solution for values of $\mu \rightarrow 0$. The small interval in which the slope of the curve is changing most rapidly is called the boundary layer. A conservative estimate of this interval is $[0, -A\mu \ln \mu]$ where A is a positive constant that is independent of μ . For the scalar equation $\mu y' = f(y, x)$ it can be shown

that $A = 2/\bar{L}$ if $-\bar{L} \leq f_y(y, x) \leq -\bar{L} < 0$. See Vasil'eva for details.

Example 4.

$$\mu y' = -y(y-1)(2x+1)10$$

with $y(0) = 2$ over the interval $[0, 1]$.

The roots of $f(y, x) = 0$ are $y = \phi(x) \equiv 0$ and $y = \phi(x) \equiv 1$. They are clearly isolated. The initial point $(2, 0)$ belongs to the domain of influence of the root $\phi(x) = 1$ because the adjointed equation

$$\frac{dy}{d\tau} = -y(y-1)10$$

with initial condition $y|_{\tau=0} = 2$ has as solution $y(\tau) = 2/(2-e^{-10\tau})$. We see at once that $\lim_{\tau \rightarrow \infty} y(\tau) = 1$. Furthermore, the root $y = \phi(x) = 1$ is positively stable in $[0,1]$ because

$$\frac{\partial f}{\partial y}(1,x) = (-2y+1)(2x+1)10 \Big|_{y=1} = -10(2x+1) < 0.$$

Therefore, we would expect that the solution $y(x,\mu) \rightarrow 1$ as $\mu \rightarrow 0$. Indeed this is the case because

$$y(x,\mu) = \frac{2}{2 - e^{-10x(x+1)/\mu}}$$

and

$$\lim_{\mu \rightarrow 0} \frac{2}{2 - e^{-10x(x+1)/\mu}} = 1 \quad \text{if } x \neq 0.$$

The paper by Vasil'eva [4] goes on to explain how to find an asymptotic expansion of the solution to the system (1.7) in terms of the small parameter μ . Her in addition to the conditions of Theorem 1.2 we assume that the righthand sides of (1.7) have continuous partial derivatives of order up to $n+2$. With this condition Vasil'eva finds an asymptotic

expansion for $0 \leq x \leq 1$ which contains n terms. Inside the boundary layer $[0, -A\mu \ln \mu]$ each term contains three functions found by solving three separate differential equations or transcendental equations. Outside the boundary layer $(-A\mu \ln \mu, 1]$ the terms are much simpler and can be determined from the variational equations. This procedure for finding the asymptotic expansion is a very tedious one and can only be explicitly calculated for the simplest problems.

It is the aim of this paper to tie together the known numerical analysis theory with the boundary layer theory in such a way that this problem can be solved with computers even as $\mu \rightarrow 0$. We do this by allowing μ to vary with the meshsize h . That is, we let $\mu = h^\gamma$ where $0 < \gamma < 1$. (Of course for each fixed value of μ , the Dahlquist theory implies convergence as $h \rightarrow 0$. But the convergence is not uniform in μ , since the derivatives of the solution $y(x, \mu)$ may become unbounded as $\mu \rightarrow 0$.)

If we attempt to apply the proof of Theorem 1.1 to this new problem, we run into serious difficulties because the proof uses the fact that

$$(1.10) \quad \lim_{\substack{h \rightarrow 0 \\ x_n = nh}} (1 + Mh)^n = e^{Mx_n}$$

where M is a positive constant. In our case, however,

the corresponding quantity that would have had to be estimated is

$$(1.11) \quad \lim_{\substack{h \rightarrow 0 \\ x_n = nh}} (1 + Mh^{1-\gamma})^n = \lim_{h \rightarrow 0} (1 + Mh^{1-\gamma})^{x_n/h} = \infty$$

as can be shown by a simple application of L'Hospital's rule: Let

$$A = \lim_{h \rightarrow 0} (1 + Mh^{1-\gamma})^{x_n/h}.$$

Then

$$\begin{aligned} \ln A &= \lim_{h \rightarrow 0} \frac{x_n \ln(1 + Mh^{1-\gamma})}{h} \\ &= \lim_{h \rightarrow 0} \frac{x_n (1 - \gamma) h^{-\gamma} M}{1 + Mh^{1-\gamma}} \quad \text{where } x_n \in (0, 1] \\ &= +\infty \end{aligned}$$

or

$$A = \infty.$$

Instead we will take advantage of the positive stability property and make use of the fact that

$$(1.12) \quad \lim_{\substack{h \rightarrow 0 \\ x_n = nh}} (1 - Mh^{1-\gamma})^n = \lim_{\substack{h \rightarrow 0 \\ x_n = nh}} (1 - Mh^{1-\gamma})^{x_n/h} = 0.$$

We pay a price for the privilege of using (1.12) instead of (1.10), namely, we must restrict ourselves to a smaller class of polynomials $\sigma(\xi)$, to guarantee convergence. This class will contain optimal method schemes of order at most $k + 1$ instead of $k + 2$ as is the case with the ODE treated in Runquist's work.

We will also use a slightly different definition of convergence for boundary layer difference equations.

DEFINITION. A difference equation associated with a boundary layer differential equation will be called convergent if for all h sufficiently small (i.e. $h \leq h_0$) there exist constants T and $\gamma < 1$ such that

$$|Y_n - y_n| \leq Th^{1-\gamma}$$

for all $nh = x_n \in [0,1]$.

We can look at this idea of relating μ to h as follows: Given a differential equation $\mu y' = f(y,x)$ we ask how can we choose h so that the associated difference equation will give an accurate approximation. If μ is sufficiently small, we choose h by the formula $h = \mu^{1/\gamma}$ where $0 < \gamma < 1$. Furthermore, we ask what is the nature of the solution to this difference equation as $\mu \rightarrow 0$. (It is obvious that if $\mu \rightarrow 0$ we must choose smaller and smaller meshwidths in order to obtain a good approximation to the solution of the differential equation.)

II. LINEAR EQUATIONS AND μ -STABILITY

We first study the numerical methods associated with the linear equation

$$(2.1) \quad \mu y' = -L_0 y + m(x)$$

on the interval $[0,1]$ with initial conditions $y(0) = 1$. L_0 is a constant which is greater than zero and $m'(x)$ is a continuous function of x . As we shall see the analysis of convergence will be independent of $m(x)$ because we are interested in only the difference $|Y_n - y(x_n)|$. We therefore consider the homogeneous equation

$$(2.2) \quad \mu y' = -L_0 y \quad \text{for } 0 \leq x \leq 1$$

with initial conditions $y(0) = 1$.

It will turn out that the results obtained for the linear equation can be generalized to the more complicated systems. Also, we will learn what restriction must be placed on the coefficients α_v and β_v ($v = 0, 1, \dots, k$) in order to insure convergence of the difference scheme

$$(2.3) \quad \sum_{i=0}^k \alpha_i Y_{n+i} = -L_0 h^{1-\gamma} \sum_{i=0}^k \beta_i Y_{n+i}$$

where we recall that $\mu = h^\gamma$, $0 < \gamma < 1$.

The method of finding a solution of (2.3) will be that of Taylor's series or small perturbations. We can prove the following theorem:

THEOREM 2.1. The solution to the difference equation

$$\sum_{i=0}^k \alpha_i Y_{n+i} = -L_0 h^{1-\gamma} \sum_{i=0}^k \beta_i Y_{n+i}$$

where L_0 is a positive constant greater than zero is

$$(2.4) \quad Y_n = C_0 \zeta_0^n + C_1 \zeta_1^n + \dots + C_{k-1} \zeta_{k-1}^n$$

where $C_0, C_1, \dots, C_{k-1}$ are constants and

$$(2.5) \quad \zeta_j = \zeta_{j0} + \left[m! \frac{\sigma(\zeta_{j0})}{\rho^{(m)}(\zeta_{j0})} (-L_0 h^{1-\gamma}) \right]^{1/m} + O(h^{\frac{2-2\gamma}{m}}).$$

ζ_{j0} is a root of $\rho(\zeta)$ of multiplicity $m \geq 1$.

Proof:

Case 1. $m = 1$. We look for a solution in the form of a power series in $h^{1-\gamma}$.

Let $Y_n = (\zeta_{j0} + qh^{1-\gamma} + \dots)^n$ where we desire to find the value of q . Substituting this expression into (2.3) gives

$$0 = \sum_{i=0}^k (\alpha_i + h^{1-\gamma} L_O \beta_i) Y_{n+i} = \sum_{i=0}^k (\alpha_i + h^{1-\gamma} L_O \beta_i) (\zeta_{j0} + q h^{1-\gamma} + \dots)^{n+i}.$$

Dividing both sides of the equation by Y_n and using the binomial theorem, we find

$$\sum_{i=0}^k (\alpha_i + h^{1-\gamma} L_O \beta_i) (\zeta_{j0}^i + i \zeta_{j0}^{i-1} q h^{1-\gamma} + \dots) = 0$$

$$\sum_{i=0}^k \alpha_i \zeta_{j0}^i + h^{1-\gamma} \sum_{i=0}^k (L_O \beta_i \zeta_{j0}^i + i \alpha_i \zeta_{j0}^{i-1} q) + \mathcal{O}(h^{2-2\gamma}) = 0.$$

The first term on the left is zero because $\rho(\zeta_{j0}) = 0$.

If we now equate coefficients of $h^{1-\gamma}$, we obtain

$$\sum_{i=0}^k (L_O \beta_i \zeta_{j0}^i + i \alpha_i \zeta_{j0}^{i-1} q) = 0$$

or

$$q = - \frac{L_O \sigma(\zeta_{j0})}{\rho'(\zeta_{j0})} \quad \text{and} \quad \rho'(\zeta_{j0}) \neq 0$$

because $m = 1$.

Case 2. $m > 1$. Here $\rho(\zeta_{j0}) = 0$ and $\rho^{(i)}(\zeta_{j0}) = 0$, $i = 1, 2, \dots, m-1$.

In this case we look for a power series expansion of the form

$$\zeta_{j_0} + qh^{(1-\gamma)/m} + O(h^{(2-2\gamma)/m})$$

$$\sum_{i=0}^k (\alpha_i + h^{1-\gamma} L_0 \beta_i) (\zeta_{j_0}^i + i \zeta_{j_0}^{i-1} qh^{(1-\gamma)/m} +$$

$$\frac{i(i-1)}{2!} \zeta_{j_0}^{i-2} q^2 h^{2(1-\gamma)/m} + \dots +$$

$$\frac{i(i-1)(i-2)\dots(i-(m-1))}{m!} \zeta_{j_0}^{i-m} q^m h^{1-\gamma} + \dots) = 0.$$

Collecting terms, we have

$$\rho(\zeta_{j_0}) + \rho'(\zeta_{j_0}) qh^{(1-\gamma)/m} + \frac{\rho''(\zeta_{j_0})}{2!} q^2 h^{2(1-\gamma)/m} +$$

$$\dots + \frac{\rho^{(m-1)}(\zeta_{j_0})}{(m-1)!} q^{m-1} h^{(m-1)(1-\gamma)/m} +$$

$$[\sigma(\zeta_{j_0}) L_0 + \frac{\rho^{(m)}(\zeta_{j_0})}{m!} q^m] h^{1-\gamma} + \dots = 0.$$

All terms but the one in brackets and those following it vanish because ζ_{j_0} has multiplicity m . Hence, equating the coefficient of $h^{1-\gamma}$ on both sides of the last equation gives

$$q^m = - \frac{m! \sigma(\zeta_{j_0}) L_0}{\rho^{(m)}(\zeta_{j_0})}$$

or

$$q = \left(- \frac{m! \sigma(\zeta_{j0}) L_0}{\rho^{(m)}(\zeta_{j0})} \right)^{1/m}.$$

That is, m distinct values of q are given by the last formula. We therefore see that if $h^{1-\gamma}$ is sufficiently small the roots ζ_j are all distinct and the theorem follows. Q.E.D.

A simple example will suffice to illustrate the fact that some well known difference methods must be excluded if we are to guarantee convergence.

Example 1. Consider the equation $\mu y' = -y$ with initial conditions $y(0) = 1$. We choose Simpson's rule formula as our difference method:

$$Y_{n+2} - Y_n = -\frac{h^{1-\gamma}}{3} (Y_{n+2} + 4Y_{n+1} + Y_n)$$

$\rho(\zeta) = \zeta^2 - 1$; the roots are $\zeta = \pm 1$. Using equation (2.5)

$$\zeta_0 = 1 - h^{1-\gamma} + \dots$$

and

$$\zeta_1 = -1 - \frac{1}{3} h^{1-\gamma} + \dots$$

The general solution is

$$Y_n = C_0(1 - h^{1-\gamma} + \dots)^n + C_1(-1 - \frac{1}{2} h^{1-\gamma} + \dots)^n.$$

However,

$$\zeta_1^n (-1)^n (1 + \frac{1}{2} h^{1-\gamma} + \dots)^n \rightarrow \infty$$

as $h \rightarrow 0$ and $x_n = nh$ by (1.11). Therefore, since we cannot insure that $C_1 = 0$ (due to rounding errors - see later discussion) Simpson's rule is divergent for boundary layer problems, when $\mu = h^\gamma$.

It is quite clear that roots of $\rho(\zeta)$ which lie inside the unit circle will not cause any problems with regard to boundedness of the solution of the difference equations. However, the simple roots on the unit circle may lead to divergent methods, as the previous example illustrates for the case $\zeta = -1$. $\zeta = 1$ is always an acceptable root because by the consistency condition ($\rho'(1) = \sigma(1)$) and by equation (2.5) we have that

$$\zeta_0 = 1 - L_0 h^{1-\gamma} + O(h^{2-2\gamma}).$$

We can easily see that the condition,

$$(2.7) \quad \frac{\sigma(-1)}{\rho'(-1)} < 0,$$

ensures that $\zeta_1^n \rightarrow 0$ as $h \rightarrow 0$ for $\zeta_{10} = -1$.

For the remaining roots on the unit circle we proceed in the following way: Let one of the roots be $\zeta_{j0} = e^{i\theta}$. Of course, another root must be given by $e^{-i\theta}$. Since we have assumed that $\sigma(\zeta)$ and $\rho(\zeta)$ have no common factors, $\sigma(e^{i\theta}) \neq 0$. Let

$$p_1 + iq_1 = - \frac{\sigma(e^{i\theta})}{\rho'(e^{i\theta})}$$

where p_1 and q_1 are real numbers. We write

$$\zeta^n = (e^{i\theta} + L_0(p_1 + iq_1)h^{1-\gamma} + \dots)^n.$$

Taking absolute values

$$\begin{aligned} |\zeta^n| &= |\zeta|^n = [\cos^2\theta + \sin^2\theta + 2L_0(p_1\cos\theta + q_1\sin\theta)h^{1-\gamma} \\ &\quad + L_0^2(p_1^2 + q_1^2)h^{2-2\gamma} + \mathcal{O}(h^{2-2\gamma})]^{n/2} \\ &= [1 + 2L_0(p_1\cos\theta + q_1\sin\theta)h^{1-\gamma} + \mathcal{O}(h^{2-2\gamma})]^{n/2}. \end{aligned}$$

Therefore, a sufficient condition for convergence in this case is

$$(2.8)' \quad p_1\cos\theta + q_1\sin\theta < 0.$$

We can state the following useful definition:

DEFINITION. A difference scheme characterized by the polynomials $\rho(\zeta)$ and $\sigma(\zeta)$ will be called μ -stable if in addition to the conditions of ordinary stability, the following conditions are satisfied for those roots of $\rho(\zeta)$ that lie on the unit circle:

1. If -1 is a root we require that

$$(2.7) \quad \frac{\sigma(-1)}{\rho'(-1)} < 0.$$

2. Let the remaining roots on the unit circle, excluding $\zeta = 1$, be denoted by $e^{i\theta_1}, e^{i\theta_2}, \dots, e^{i\theta_\ell}$.
Let

$$p_j + iq_j = - \frac{\sigma(e^{i\theta_j})}{\rho'(e^{i\theta_j})} \quad j = 1, 2, \dots, \ell.$$

We require that

$$(2.8) \quad p_j \cos \theta_j + q_j \sin \theta_j < 0 \quad j = 1, 2, \dots, \ell.$$

We immediately notice that conditions (2.8) really only contain $\ell/2$ conditions because we are dealing with complex conjugates. That is, if $e^{i\theta_1}$ is a root so is $e^{-i\theta_1}$. But

$$p_1 + iq_1 = - \frac{\sigma(e^{i\theta_1})}{\rho'(e^{i\theta_1})}$$

and

$$\begin{aligned} p_2 + iq_2 &= - \frac{\sigma(e^{i\theta_2})}{\rho'(e^{i\theta_2})} = - \frac{\sigma(e^{-i\theta_1})}{\rho'(e^{-i\theta_1})} \\ &= p_1 - iq_1 \end{aligned}$$

where we have let $\theta_2 = -\theta_1$. Hence,

$$p_1 = p_2 \quad \text{and} \quad q_1 = -q_2.$$

Now,

$$p_2 \cos \theta_2 + q_2 \sin \theta_2 < 0$$

is equivalent to

$$p_1 \cos(-\theta_1) - q_1 \sin(-\theta_1) < 0$$

or

$$p_1 \cos \theta_1 + q_1 \sin \theta_1 < 0.$$

Therefore, the two conditions are equivalent.

The conditions of μ -stability can be thought of as merely conditions on the coefficients, β_v , $v = 0, 1, \dots, k$. That is, given a polynomial $\rho(\zeta)$ that satisfies ordinary stability, we attempt to find a related $\sigma(\zeta)$ that satisfies the consistency conditions and the μ -stability conditions. A few such examples will clarify this point.

Example 2. Let $\rho(\zeta) = \zeta^2 - 1$; the roots are $\zeta = \pm 1$. The general form of $\sigma(\zeta)$ is $\sigma(\zeta) = \beta_2 \zeta^2 + \beta_1 \zeta + \beta_0$. By consistency

$$(2.9) \quad \rho'(1) = 2 = \sigma(1) = \beta_2 + \beta_1 + \beta_0.$$

By condition (2.7) we have

$$\frac{\sigma(-1)}{\rho'(-1)} = \frac{\beta_2 - \beta_1 + \beta_0}{-2} < 0$$

or

$$(2.10) \quad \beta_2 - \beta_1 + \beta_0 > 0.$$

Combining (2.9) and (2.10) gives

$$\beta_1 - \beta_0 = \beta_1 + \beta_0 >$$

or

$$(2.11) \quad \beta_1 < 1.$$

Thus the inequality (2.11) is equivalent to the condition that the polynomials $\rho(\zeta) = \zeta^2 - 1$ and $\sigma(\zeta) = \beta_2 \zeta^2 + \beta_1 \zeta + \beta_0$ satisfy the condition of μ -stability.

We now see why Simpson's rule was divergent. In Simpson's rule $\beta_1 = \frac{4}{3}$, which does not satisfy (2.11).

Example 3. Let $\rho(\zeta) = \zeta^4 - 1$. The roots are $\zeta = \pm 1$ and $\zeta = \pm i$. $\sigma(\zeta) = \beta_4 \zeta^4 + \beta_3 \zeta^3 + \beta_2 \zeta^2 + \beta_1 \zeta + \beta_0$. By consistency,

$$(2.11)' \quad 4 = \rho'(1) = \sigma(1) = \beta_4 + \beta_3 + \beta_2 + \beta_1 + \beta_0.$$

Condition (2.7) implies

$$(2.12) \quad \frac{\sigma(-1)}{\rho'(-1)} = -\frac{1}{4} (\beta_4 - \beta_3 + \beta_2 - \beta_1 + \beta_0) < 0.$$

Combining (2.11)' and (2.12) gives

$$(2.13) \quad 2 > \beta_3 + \beta_1.$$

For the root $\zeta = i$ we have

$$p_1 + iq_1 = -\frac{\sigma(i)}{\rho'(i)} = -\frac{1}{4} [(\beta_4 - \beta_2 + \beta_0)i + (\beta_3 - \beta_1)]$$

or

$$q_1 = -\frac{1}{4} (\beta_4 - \beta_2 + \beta_0) \quad p_1 = -\frac{1}{4} (\beta_3 - \beta_1).$$

We require by (2.8) that

$$q_1 < 0$$

or

$$\beta_4 - \beta_2 + \beta_0 > 0.$$

Combining the last expression with (2.11) yields

$$(2.14) \quad 4 > \beta_3 + 2\beta_2 + \beta_1.$$

Equations (2.13) and (2.14) express μ -stability for this example. An example of a convergent method would be

$$\rho(\zeta) = \zeta^4 - 1$$

$$\sigma(\zeta) = \zeta^4 + \frac{1}{2} \zeta^3 + \zeta^2 + \frac{1}{2} \zeta + 1.$$

Check: (2.13) is satisfied because

$$2 > \frac{1}{2} + \frac{1}{2};$$

(2.14) is satisfied because

$$4 > \frac{1}{2} + 2(1) + \frac{1}{2}.$$

It will be recalled that in the ordinary differential equation theory stability of the consistent difference scheme, i.e. boundedness of its solution was a necessary and sufficient condition for convergence. The case with boundary layer equations is a bit different; that is, μ -stability isn't in general a necessary condition for convergence. This follows because we have not considered the cases when

$$(2.15) \quad p_j \cos \theta_j + q_j \sin \theta_j = 0.$$

Here the precise determination of a necessary and sufficient condition seems much more difficult and we leave this as an open problem.

In order to be complete we should give a simple example where

$$p_j \cos \theta_j + q_j \sin \theta_j = 0$$

but $\sigma(e^{i\theta_j}) \neq 0$, since we have assumed that $\sigma(\zeta)$ and $\rho(\zeta)$ have no common factors.

Example 4. Consider $\rho(\zeta) = \zeta^3 - \zeta^2 + \zeta - 1$. The roots are $\zeta = 1, \pm i$. $\rho'(1) = 2$. Therefore, $2 = \beta_3 + \beta_2 + \beta_1 + \beta_0$ and the μ -stability condition implies

$$1 > \beta_1 + \beta_2.$$

Suppose instead, we let

$$\beta_1 + \beta_2 = 1.$$

Let $\beta_1 = 1/4$ and $\beta_2 = 3/4$. Also, let $\beta_0 = \beta_3 = 1/2$.

Then $\sigma(\zeta) = \frac{1}{2} \zeta^3 + \frac{3}{4} \zeta^2 + \frac{1}{4} \zeta + \frac{1}{2}$ and

$$\sigma(1) = 2 \neq 0$$

$$\sigma(i) = -\frac{1}{4} - \frac{1}{4}i \neq 0, \quad \rho'(i) = -2 - 2i; \quad p = -\frac{1}{8}, \quad q = 0$$

$$\sigma(-i) = -\frac{1}{4} + \frac{1}{4}i \neq 0, \quad \rho'(-i) = -2 + 2i; \quad p = -\frac{1}{8}, \quad q = 0.$$

Not only can examples be found which allow $p_j \cos \theta_j + q_j \sin \theta_j$ to be equal to zero, but we can also give examples in which $p_j \cos \theta_j + q_j \sin \theta_j < 1/m$ where m is arbitrarily large. In practice such a scheme should be avoided because the convergence of the scheme would require a very small value of h .

Example 5. Let $\rho(\zeta) = \zeta^2 - 1$. Then $\sigma(1) = \beta_2 + \beta_1 + \beta_0 = 2$ and the μ -stability condition implies $\beta_1 < 1$. We choose

$$\sigma(-1) = \beta_2 - \beta_1 + \beta_0 = 1/m$$

$$\sigma(1) + \sigma(-1) = 2\beta_2 + 2\beta_0 = 2 + \frac{1}{m}.$$

$$\text{Let } \beta_0 = 1. \text{ Then } \beta_2 = \frac{1}{2m} \text{ and } \beta_1 = 1 - \frac{1}{2m}.$$

Therefore

$$\begin{aligned} \zeta_1 &= -1 - \frac{\sigma(-1)}{\rho'(-1)} h^{1-\gamma} + O(h^{2(1-\gamma)}) \\ &= -1 + \frac{1}{2m} h^{1-\gamma} + O(h^{2(1-\gamma)}) \end{aligned}$$

is a root of $\rho(\zeta) + h^{1-\gamma}\sigma(\zeta) = 0$.

Because of this last example it seems wise also to consider a stronger kind of stability.

DEFINITION. A difference scheme characterized by the polynomials $\rho(\zeta)$ and $\sigma(\zeta)$ will be called relatively stable if the roots of $\rho(\zeta) + h^{1-\gamma}\sigma(\zeta) = 0$ have the property that

$$|\zeta_i| \leq |\zeta_0| = 1 - h^{1-\gamma} + O(h^{2(1-\gamma)}) \quad i = 1, 2, \dots, k-1.$$

For a relatively stable scheme we find that

$$\frac{\sigma(-1)}{\rho'(-1)} < -1$$

and

$$p_j \cos \theta_j + q_j \sin \theta_j < -\frac{1}{2}.$$

These conditions are more stringent than the inequalities (2.7) and (2.8) which are required for μ -stability.

The proof that μ -stability implies convergence for the linear difference scheme (2.3) requires a few lemmas. Instead of using subscripts for each new constant that is introduced, we will use the notation that C equals a constant. When we have inequalities like

$$(C^2 + 5) \leq C$$

we will mean that a new constant C_1 exists such that

$$C^2 + 5 \leq C_1.$$

Of course, the fact that these constants are independent of h for $0 < h \leq h_0$ is to be demonstrated for some h_0 .

LEMMA 1. Let the consistent difference equation

$$\sum_{i=0}^k \alpha_i Y_{n+i} = -L_0 h^{1-\gamma} \sum_{i=0}^k \beta_i Y_{n+i}$$

be μ -stable. L_0 is a positive constant with the property that $-\bar{L} \leq -L_0 \leq -\bar{L} < 0$. Let $\bar{\Phi}_i$ be the solution of this difference equation with initial conditions $\bar{\Phi}_0 = \bar{\Phi}_1 = \dots = \bar{\Phi}_{k-2} = 0$, $\bar{\Phi}_{k-1} = (\alpha_k + h^{1-\gamma} L_0 \beta_k)^{-1}$.

Then, for all n ,

$$(2.16) \quad \sum_{i=0}^{n-1} |\Phi_i| \leq \frac{C}{h^{1-\gamma}}$$

for h sufficiently small and where C , a constant, may be chosen independent of L_0 , i.e. C is a uniform bound for all L_0 such that $-\bar{L} \leq -L_0 \leq -\bar{L} < 0$.

Proof: The solution to the difference equation is

$$\Phi_i = C_0(h)\zeta_0^i + C_1(h)\zeta_1^i + \cdots + C_{k-1}(h)\zeta_{k-1}^i$$

where $\zeta_0, \zeta_1, \dots, \zeta_{k-1}$ are linearly independent solutions of the difference equation and are distinct if h is small enough. ζ_j was given earlier as

$$\begin{aligned} \zeta_j &= \zeta_{j0} + (m! \frac{\delta(\zeta_{j0})}{\rho^{(m)}(\zeta_{j0})} (-L_0 h^{1-\gamma}))^{1/m} \\ &\quad + \mathcal{O}(h^{2(1-\gamma)/m}) \end{aligned}$$

where $\rho(\zeta_{j0}) = 0$ and m is the multiplicity of ζ_{j0} .

By Cramer's rule, we can write

$$c_j(h) = \frac{\begin{vmatrix} 1 & 1 & \dots & 1 & 0 & & 1 & \dots & 1 \\ \zeta_0 & \zeta_1 & & \zeta_{j-1} & 0 & & \zeta_{j+1} & & \zeta_{k-1} \\ \zeta_0^2 & \zeta_1^2 & & \zeta_{j-1}^2 & 0 & & \zeta_{j+1}^2 & & \zeta_{k-1}^2 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ \zeta_0^{k-1} & \zeta_1^{k-1} & \dots & \zeta_{j-1}^{k-1} & (\alpha_k + h^{1-\gamma_{L_0}} \beta_k)^{-1} & & \zeta_{j+1}^{k-1} & \dots & \zeta_{k-1}^{k-1} \end{vmatrix}}{\begin{vmatrix} 1 & 1 & & & & & 1 \\ \zeta_0 & \zeta_1 & & & & & \zeta_{k-1} \\ \zeta_0^2 & \zeta_1^2 & & & & & \zeta_{k-1}^2 \\ \vdots & \vdots & & & & & \vdots \\ \zeta_0^{k-1} & \zeta_1^{k-1} & & & & & \zeta_{k-1}^{k-1} \end{vmatrix}}$$

The numerator we denote by D_j , the denominator by W (the Wronskian). We observe that D_j and W reduce to Vandermonde's determinant.

$$W = \prod_{j < i} (\zeta_i - \zeta_j)$$

$$|D_j| = |(\alpha_k + h^{1-\gamma_{L_0}} \beta_k)^{-1}| \prod_{\substack{t < s \\ s \neq j \\ t \neq j}} |\zeta_s - \zeta_t|.$$

Therefore,

$$|c_j(h)| = \frac{|(\alpha_k + h^{1-\gamma} L_0 \beta_k)^{-1}|}{\prod_{\substack{i=0 \\ i \neq j}}^{k-1} |\zeta_i - \zeta_j|}.$$

A power of h is the leading term of a difference expression $\zeta_i - \zeta_j$ only when $\zeta_{i0} = \zeta_{j0}$. Let us assume that $\zeta_{00} = 1$ and ζ_{j0} has multiplicity m_j .

Thus

$$|c_j(h)| \leq \frac{C}{h^{[(1-\gamma)/m_j](m_j-1)}}.$$

If $m_j = 1$ (ζ_{j0} is a simple root), then by μ -stability there exists a constant $L > 0$ such that $|\zeta_j|^2 \leq 1 - Lh^{1-\gamma}$. In this case

$$\begin{aligned} |c_j(h)| \sum_{i=0}^{n-1} |\zeta_j|^i &\leq C \sum_{i=0}^{n-1} (1 - Lh^{1-\gamma})^{i/2} \\ &\leq C \left[\frac{1 - (1 - Lh^{1-\gamma})^{n/2}}{1 - (1 - Lh^{1-\gamma})^{1/2}} \right] \\ &\leq \frac{C}{h^{1-\gamma}} \end{aligned}$$

for h sufficiently small. If $m_j > 1$, then by μ -stability for h sufficiently small $|\zeta_j| \leq r < 1$. Here

$$\begin{aligned}
|c_j(h)| \sum_{i=0}^{n-1} |\zeta_j|^i &\leq \frac{C}{h^{(1-\gamma)(m_j-1)/m_j}} \sum_{i=0}^{n-1} r^i \\
&\leq \frac{C}{h^{1-\gamma}} \frac{1-r^n}{1-r} \\
&\leq \frac{C}{h^{1-\gamma}}.
\end{aligned}$$

Combining these results gives (2.16). Q.E.D.

Remark: If we allow γ to be equal to zero and L_0 to be a positive or negative constant, then (2.16) becomes

$$\sum_{i=0}^{n-1} |\Phi_i| \leq \frac{C}{h} \quad 0 \leq x_n \leq 1 \quad \text{and} \quad \gamma = 0.$$

The proof is similar to that of Lemma 1 with the use of

$$\lim_{\substack{h \rightarrow 0 \\ x_n = nh}} (1 \pm Mh)^n = e^{\pm Mx_n} \leq C$$

instead of (1.12).

Using the same conditions as in Lemma 1, we now prove
LEMMA 2.

$$|\Phi_n| \leq \Phi \left(\frac{r^n}{h^{(1-\gamma)(m-1)/m}} + (1 - Lh^{1-\gamma})^n \right)$$

for h sufficiently small, where m equals the maximum multiplicity of the roots ζ_{j0} and $r = \frac{1}{2}(1 + \max_{|\zeta_{j0}| < 1} |\zeta_{j0}|) < 1$.

L and Φ are positive constants independent of the particular L_0 chosen.

Proof: In the course of the proof of Lemma 1, it was shown that

$$\Phi_n = c_0 \zeta_0^n + c_1 \zeta_1^n + \dots + c_{k-1} \zeta_{k-1}^n$$

where

$$|c_j| \leq \frac{c}{h^{(1-\gamma)(m_j-1)/m_j}} \quad \text{if } |\zeta_{j0}| < 1$$

$$|c_j| \leq c \quad \text{if } |\zeta_{j0}| = 1.$$

Also, by μ -stability for h sufficiently small

$$|\zeta_j| \leq r \quad \text{if } |\zeta_{j0}| < 1$$

$$|\zeta_j| \leq 1 - Lh^{1-\gamma} \quad \text{if } |\zeta_{j0}| = 1$$

where $L > 0$.

From these inequalities the lemma easily follows.

We complete this chapter by proving the following theorem about the difference method associated with the

equation (2.1); i.e. $\mu y' = -L_0 y + m(x)$.

THEOREM 2.2. Let the consistent finite difference equation

$$(2.17) \quad \sum_{i=0}^k \alpha_i Y_{n+i} = h^{1-\gamma} \sum_{i=0}^k \beta_i (-L_0 Y_{n+i} + m(x_{n+i})) + R_n$$

where R_n is the round-off error, satisfy the following conditions:

- a) $R_n = O(h^{2-2\gamma})$ as $h \rightarrow 0$;
- b) The finite difference equation is μ -stable;
- c) $m'(x)$ is continuous for $0 \leq x \leq 1$;
- d) $|Y_i - y(x_i)| \leq Th^{1-\gamma}$ for $i = 0, 1, \dots, k-1$.

Then the difference equation (2.17) is convergent, i.e.

$$|Y_n - y(x_n)| \leq Ch^{1-\gamma} \quad \text{for } n = 0, 1, 2, \dots \quad \text{and } 0 \leq x_n \leq 1;$$

C is independent of h .

Proof: The proof is divided into two parts. We first prove $e_n = Y_n - y(x_n) = O(h^{1-\gamma})$ for $0 \leq n \leq N_0 = \frac{\ln h^{1-\gamma}}{\ln r}$ where r was defined in Lemma 2 as $r = \frac{1}{2}(1 + \max |\zeta_{j_0}|)$. In the second part we prove $e_n = O(h^{1-\gamma})$ for $|\zeta_{j_0}| < 1$

$N_0 \leq n \leq N$ where $x_N = 1$.

First we note that

$$\mu y'' = -L_0 y' + m'(x) = -L_0 \left(\frac{-L_0 y + m(x)}{\mu} \right) + m'(x)$$

or

$$\begin{aligned} h^2 y'' &= h^{2(1-\gamma)} (L_0^2 y - L_0 m(x)) + h^{2-\gamma} m'(x) \\ &= \mathcal{O}(h^{2(1-\gamma)}) \end{aligned}$$

where we have used statement c) of the hypothesis.

The exact solution $y(x)$ of (2.1) will now satisfy the difference equation

$$(2.18) \quad \sum_{i=0}^k \alpha_i y_{n+i} = h^{1-\gamma} \sum_{i=0}^k \beta_i (-L_0 y_{n+i} + m(x_{n+i})) + T_n$$

where $y_n = y(x_n)$ and $T_n = \mathcal{O}(h^{2(1-\gamma)})$.

Subtracting (2.18) from (2.17) and letting $e_i = Y_i - y_i$, we obtain

$$(2.19) \quad \sum_{i=0}^k \alpha_i e_{n+i} = -h^{1-\gamma} L_0 \sum_{i=0}^k \beta_i e_{n+i} + \mathcal{O}(h^{2(1-\gamma)}).$$

Let the right hand side of (2.19) be denoted by Q_n .

The solution to this difference equation is given in Hull and Luxemburg [3] as

$$(2.20) \quad e_n = \begin{cases} \sum_{i=0}^{n-k} g_{n-i-1} Q_i + \theta_n & n \geq k \\ \theta_n & n < k \end{cases}$$

where g_n is defined by

$$\sum_{i=0}^k \alpha_i g_{n+i} = 0$$

and $g_0 = g_1 = \dots = g_{k-2} = 0$, $g_{k-1} = \alpha_k^{-1}$; and where

$$\theta_n = \sum_{i=0}^{k-1} \left(\sum_{j=0}^{k-i-1} \alpha_{k-j} g_{n+k-i-j-1} \right) e_i \quad \text{for } n \geq k.$$

Consequently, setting $\max |g_n| = g < \infty$, by μ -stability

$$|\theta_n| \leq kag \sum_{i=0}^{k-1} |e_i| \leq k^2 agTh^{1-\gamma} \leq K_1 h^{1-\gamma}$$

by condition d).

Here $\alpha = \max |\alpha_i|$ and $K_1 \equiv k^2 agT$.

Letting $\beta = \max |\beta_i|$ we obtain from (2.19) and (2.20)

$$\begin{aligned} |e_n| &\leq h^{1-\gamma} \beta g L_0 |e_n| + h^{1-\gamma} \beta g L_0 (k+1) \sum_{i=0}^{n-1} |e_i| \\ &\quad + g \sum_{i=k}^n O(h^{2(1-\gamma)}) + |\theta_n|. \end{aligned}$$

Now if $h \leq h_0$ where $h_0^{1-\gamma} \beta g L_0 < 1$ and $0 \leq n \leq N_0 =$

$\frac{\ln h^{1-\gamma}}{\ln r}$, then

$$(2.21) \quad |e_n| \leq h^{1-\gamma} A \sum_{i=0}^{n-1} |e_i| + K_2 h^{1-\gamma}$$

where

$$A = \frac{\beta g L_0 (k+1)}{1-h^{1-\gamma} \beta g L_0} \quad \text{and} \quad K_3 = \frac{g K_2 N_0 h^{1-\gamma} + K_1}{1-h_0^{1-\gamma} \beta g L_0}$$

where K_2 is a bound for R_n and T_n .

From (2.21) and a simple induction (see Henrici [2] p. 244) one can show

$$\begin{aligned} |e_n| &\leq K_3 h^{1-\gamma} (1 + A h^{1-\gamma})^n \\ &\leq K_3 h^{1-\gamma} e^{n h^{1-\gamma} A} \\ (2.22) \quad &\leq K_3 h^{1-\gamma} e^{A(h^{1-\gamma} \ln h^{1-\gamma}) / \ln r} \\ &\leq O(h^{1-\gamma}) \quad \text{for } 0 \leq n \leq N_0. \end{aligned}$$

For $N_0 \leq n \leq N$ we consider the difference equation

$$\sum_{i=0}^k (\alpha_i + h^{1-\gamma} L_0 \beta_i) e_{n+i} = R_n - T_n$$

with the solution in the form

$$(2.23) \quad e_n = \begin{cases} \sum_{i=0}^{n-k} \Phi_{n-i-1} (R_i - T_i) + \Psi_n & n \geq k \\ \Psi_n & n < k \end{cases}$$

where Φ_n satisfies the difference equation

$$\sum_{i=0}^k (\alpha_i + h^{1-\gamma} L_0 \beta_i) \Phi_{n+i} = 0$$

with initial conditions

$$\Phi_0 = \Phi_1 = \dots = \Phi_{k-2} = 0, \quad \Phi_{k-1} = (\alpha_k + h^{1-\gamma} L_0 \beta_k)^{-1};$$

$$\Psi_n = \sum_{i=0}^{k-1} \left(\sum_{j=0}^{k-i-1} (\alpha_{k-j} + h^{1-\gamma} \beta_{k-j} L_0) \right) \Phi_{n+k-i-j-1} e_i$$

for $n \geq k$.

Since here we are interested in the case $n \geq N_0$, we can write, using Lemma 2 and condition d)

$$|\Psi_n| \leq (\alpha + h^{1-\gamma} \beta L_0) (|\Phi_{n+k-1}| + \dots + |\Phi_n|) \sum_{i=0}^{k-1} |e_i|$$

$$\leq k^2 (\alpha + h^{1-\gamma} \beta L_0) \Phi \left(\frac{r^{N_0}}{h^{1-\gamma}} + (1 - L h^{1-\gamma})^{N_0} \right) T h^{1-\gamma}$$

$$= O(h^{1-\gamma}) \quad \text{for } N_0 \leq n$$

because $r^{N_0} = e^{N_0 \ln r} = e^{\ln h^{1-\gamma} N_0} = h^{1-\gamma}$.

By Lemma 1 it follows that

$$\begin{aligned}
\sum_{i=0}^{n-k} |\Phi_{n-i-1}| |R_i - T_i| &\leq \sum_{i=0}^{n-k} |\Phi_{n-i-1}| \mathcal{O}(h^{2(1-\gamma)}) \\
&= \mathcal{O}(h^{1-\gamma}).
\end{aligned}$$

Combining these results with (2.23) gives

$$(2.24) \quad e_n = \mathcal{O}(h^{1-\gamma}) \quad \text{for } N_0 \leq n \leq N.$$

Q.E.D.

III.

OPTIMAL METHODS

The criterion of μ -stability leads to some interesting results with regard to optimal methods. By optimal methods we mean given a polynomial $\rho(\zeta)$ of degree k and satisfying $\rho(1) = 0$ what is the maximum order that can be achieved in the operator $L[y(x); h^{1-\gamma}]$ by choosing a suitable polynomial $\sigma(\zeta)$. For the ODE case the order of a stable operator when k is odd cannot exceed $k + 1$ and when k is even cannot exceed $k + 2$. We now briefly describe the theory of optimal methods as given by Henrici [2].

We associate with the polynomials $\rho(\zeta)$ and $\sigma(\zeta)$ the function

$$(3.1) \quad \phi(\zeta) = \frac{\rho(\zeta)}{\ln \zeta} - \sigma(\zeta)$$

where $\ln \zeta$ is made single-valued by cutting the complex plane along the negative real axis and setting $\ln 1 = 0$. Using (3.1) and a few simple properties of analytic functions, we can prove the following:

THEOREM 3.1. The difference operator $L[y(x); h]$ associated with the polynomials $\rho(\zeta)$ and $\sigma(\zeta)$ has the exact order p if and only if the function $\phi(\zeta)$

has a zero of the exact order p at $\zeta = 1$.

Proof: See Henrici [2].

THEOREM 3.2. Let $\rho(\zeta)$ be a polynomial of degree k satisfying $\rho(1) = 0$ and let k' be an integer, $0 \leq k' \leq k$. Then there exists a unique polynomial $\sigma(\zeta)$ of degree $\leq k'$ such that the order of the associated difference operator is at least $k'+1$.

Proof: See Henrici [2].

$\frac{\rho(\zeta)}{\ln \zeta}$ is analytic at $\zeta = 1$, so suppose we let

$$\frac{\rho(\zeta)}{\ln \zeta} = c_0 + c_1(\zeta-1) + c_2(\zeta-1)^2 + \dots$$

We choose

$$\sigma(\zeta) = c_0 + c_1(\zeta-1) + \dots + c_{k'}(\zeta-1)^{k'}.$$

By the theorems we have that

$$\phi(\zeta) = c_{p+1}(\zeta-1)^p + \mathcal{O}((\zeta-1)^{p+1}).$$

An example will illustrate the technique described :

Example 1. Let $\rho(\zeta) = (\zeta-1)(\zeta-\lambda)$ where λ is real and $-1 \leq \lambda < 1$.

$$\begin{aligned} \frac{\rho(\zeta)}{\ln \zeta} &= \frac{(\zeta-1)[1-\lambda + (\zeta-1)]}{\ln[1 + (\zeta-1)]} \\ &= 1-\lambda + \frac{3-\lambda}{2}(\zeta-1) + \frac{5+\lambda}{12}(\zeta-1)^2 - \frac{(1+\lambda)}{24}(\zeta-1)^3 + \mathcal{O}((\zeta-1)^4). \end{aligned}$$

Therefore,

$$\begin{aligned} \sigma(\zeta) &= 1-\lambda + \frac{3-\lambda}{2}(\zeta-1) + \frac{5+\lambda}{12}(\zeta-1)^2 \\ &= -\frac{1+5\lambda}{12} + \frac{2-2\lambda}{3}\zeta + \frac{5+\lambda}{12}\zeta^2. \end{aligned}$$

Note that for $\lambda \neq -1$ the order is 3 and for $\lambda = -1$ the order is 4. Thus, in the standard ODE case it is sometimes possible to achieve $p = k + 2$. We investigate this further.

We introduce a new variable

$$(3.2) \quad z = \frac{\zeta-1}{\zeta+1} \quad \text{or} \quad \zeta = \frac{1+z}{1-z}$$

and two new polynomials of degree $\leq k$

$$(3.3) \quad r(z) = \left(\frac{1-z}{2}\right)^k \rho\left(\frac{1+z}{1-z}\right)$$

$$(3.4) \quad s(z) = \left(\frac{1-z}{2}\right)^k \sigma\left(\frac{1+z}{1-z}\right).$$

Henrici [2] shows that (3.3) can be written

$$(3.5) \quad r(z) = a_1 z + a_2 z^2 + \dots + a_{k-1} z^{k-1} + a_k z^k$$

where the coefficients are real and

$$(3.6) \quad a_1 > 0$$

$$(3.7) \quad a_\mu \geq 0 \quad \mu = 2, 3, \dots, k.$$

We now consider the function

$$(3.8) \quad p(z) = \left(\frac{1-z}{2}\right)^k \phi\left(\frac{1+z}{1-z}\right) = \frac{r(z)}{\ln \frac{1+z}{1-z}} - s(z).$$

The function $p(z)$ has a zero of order p at $z = 0$ if and only if $\phi(\xi)$ has a zero of order p at $\xi = 1$ and thus, by Theorem 3.1, if and only if the operator associated with $\rho(\xi)$ and $\sigma(\xi)$ has order p . Consequently, if the operator has order p , then

$$(3.9) \quad s(z) = b_0 + b_1 z + b_2 z^2 + \dots + b_{p-1} z^{p-1}$$

where

$$(3.10) \quad \frac{z r(z)}{\left[\ln \frac{1+z}{1-z}\right] z} = b_0 + b_1 z + b_2 z^2 + \dots .$$

The degree of $s(z)$ cannot exceed k ; therefore, a stable operator with $p > k + 1$ exists only if $b_{k+1} = \dots = b_{p-1} = 0$. We use the following property of the $\ln \zeta$

$$(3.11) \quad \frac{z}{\ln \frac{1+z}{1-z}} = C_0 + C_2 z^2 + C_4 z^4 + \dots$$

where

$$(3.12) \quad C_{2v} < 0, \quad v = 1, 2, \dots$$

If k is odd, then

$$(3.13) \quad b_{k+1} = C_2 a_k + C_4 a_{k-2} + \dots + C_{k+1} a_1.$$

Using (3.6), (3.7) and (3.12), it follows that $b_{k+1} < 0$. Hence,

THEOREM 3.3. If k is odd, the order of a stable operator cannot exceed $k + 1$.

If on the other hand k is even, then

$$(3.14) \quad b_{k+1} = C_2 a_k + C_4 a_{k-2} + \dots + C_k a_2.$$

A necessary and sufficient condition for $b_{k+1} = 0$ is that $a_2 = a_4 = \dots = a_k = 0$. That is, $r(-z) = -r(z)$.

By stability the real parts of the roots of $r(z)$ must be ≤ 0 . Since $r(z)$ is an odd function here, $r(z)$ cannot have roots with negative real part. We, therefore, have by (3.2) that all roots of $\rho(\zeta)$ lie on $|\zeta| = 1$. Since $a_k = 0$, the degree of $r(z)$ is $k - 1$. Thus -1 is a root of $\rho(\zeta)$ (otherwise $r(z)$ would be of degree k). Using (3.6), (3.7) and (3.12) again

$$(3.15) \quad b_{k+2} = c_4 a_{k-1} + c_6 a_{k-3} + \dots + c_{k+2} a_1$$

is negative, we find that the order of a stable ODE operator cannot exceed $k + 2$.

THEOREM 3.4. A necessary and sufficient condition for $p = k + 2$ is that k be even, that all roots of $\rho(\zeta)$ have modulus 1, and that $\sigma(\zeta)$ be determined by (3.5).

We are now prepared to impose the condition of μ -stability and investigate the changes in the above theory. It is quite clear that if $\zeta_{00} = 1$ and $|\zeta_{j0}| < 1$ ($j = 1, 2, \dots, k-1$), then we can use formula (3.1) to obtain a difference operator of order $p = k + 1$. We also have the following result:

THEOREM 3.5. All stable methods of order $p = k + 2$ are not μ -stable.

Proof: Since $\zeta = -1$ must be a root of $\rho(\zeta)$ in order that $p = k + 2$, it will suffice to show that $\frac{\sigma(-1)}{\rho'(-1)} < 0$ cannot be satisfied. Since $\rho(\zeta)$ has the form

$$\rho(\zeta) = (\zeta-1)(\zeta+1)(\zeta-e^{i\theta_1})(\zeta-e^{-i\theta_1}) \dots (\zeta-e^{\frac{i\theta_{k-2}}{2}})(\zeta-e^{-\frac{i\theta_{k-2}}{2}})$$

$$\rho'(-1) = -2(-1-e^{i\theta_1})(-1-e^{-i\theta_1}) \dots (-1-e^{\frac{i\theta_{k-2}}{2}})(-1-e^{-\frac{i\theta_{k-2}}{2}})$$

$$< 0.$$

Therefore, we need only show that $\sigma(-1) < 0$. Since by (3.2) $\zeta = -1$ corresponds to $z = \infty$, we have, using (3.4)

$$\begin{aligned} \sigma(-1) &= 2^k \lim_{z \rightarrow \infty} \frac{s(z)}{(1-z)^k} \\ &= 2^k \lim_{z \rightarrow \infty} \frac{b_0 + b_1 z + \dots + b_k z^k}{(1-z)^k} \\ &= \frac{2^k k! b_k}{(-1)^k k!} \\ &= 2^k b_k \quad \text{by L'Hospital's rule.} \end{aligned}$$

Since k is even,

$$b_k = c_2 a_{k-1} + c_4 a_{k-3} + \dots + c_k a_1.$$

Using $a_1 > 0$, $a_v \geq 0$ ($v = 2, 3, \dots$), and $c_{2v} < 0$, we obtain $b_k < 0$ and therefore $\sigma(-1) < 0$.

Hence, $\frac{\sigma(-1)}{\rho'(-1)} > 0$ and the result follows.

Immediate consequences of this theorem are the following corollaries:

COROLLARY 1. If k is even and $\rho(-1) = 0$, then the unique stable scheme of order $p = k + 1$ is not μ -stable.

Proof: Since k is even and $\zeta = -1$ is a root, there exist an even number of roots of $\rho(\zeta)$ which are real. Denote them by $1, -1, \lambda_1, \lambda_2, \dots, \lambda_{2s}$. As in the theorem, we have

$$\rho'(-1) = -2(-1-\lambda_1) \dots (-1-\lambda_{2s}) \quad (\text{positive number}).$$

By virtue of stability $|\lambda_i| < 1$. Thus $\rho'(-1) < 0$. As in the proof of the theorem $\sigma(-1) < 0$. Thus, the scheme is not μ -stable.

COROLLARY 2. If k is odd and the roots of $\rho(\zeta)$ are $\lambda_1 = 1, \lambda_2 = -1, \lambda_3, \dots, \lambda_k$ where $|\lambda_i| < 1$ for $i = 3, 4, \dots, k$. Then the unique stable scheme of order $p = k + 1$ is μ -stable.

Proof: By hypothesis it easily follows that $\rho'(-1) > 0$. Since k is odd

$$b_k = c_2 a_{k-1} + c_4 a_{k-3} + \dots + c_{k-1} a_2.$$

Using

$$c_{2v} < 0 \quad v = 1, 2, \dots$$

and

$$a_\mu \geq 0 \quad \mu = 1, 2, \dots,$$

we obtain $b_k \leq 0$. But $b_k = 0$ is equivalent to $\sigma(-1) = 0$ ($\sigma(-1) = 2^k b_k$) and we have excluded the case of common factors of $\sigma(\zeta)$ and $\rho(\zeta)$. Therefore, $\sigma(-1) < 0$ and

$$\frac{\sigma(-1)}{\rho(-1)} < 0. \quad \text{Q.E.D.}$$

In concluding we say that the best precision that we can hope for is $p = k + 1$. From Corollary 1 we see that this value of p is not always obtainable. In general when $\rho(\zeta)$ has roots on the unit circle and k is odd, we must test separately whether each complex root on the unit circle satisfies the conditions of μ -stability.

Example 2. Let $\rho(\zeta) = \zeta^3 - \zeta^2 + \zeta - 1$. The roots are $1, \pm i$, and the μ -stability condition reduces to $1 > \beta_1 + \beta_2$. The unique optimal method for this particular choice of $\rho(\zeta)$ is characterized by

$$\sigma(\zeta) = \frac{1}{12}(5\zeta^3 + 7\zeta^2 + 7\zeta + 5)$$

where

$$\beta_1 + \beta_2 = \frac{7}{12} + \frac{7}{12} = \frac{7}{6} \not< 1.$$

Therefore, this optimal method is not μ -stable.

IV. BOUNDARY LAYER EQUATIONS OF THE FIRST ORDER IN ONE UNKNOWN

In this chapter we discuss the numerical treatment of the equation (1.6) $\mu y' = f(y, x)$ over the interval $[0, 1]$ with initial conditions $y(0) = a$. We will assume a slightly more general condition of positive stability than the one given in Chapter 1. In addition we require that the real parts of the roots of the characteristic equation

$$\text{Det} \left(\frac{\partial f}{\partial y} (y(x, \mu), z(x, \mu), x) - \lambda I \right) = 0$$

are negative and bounded away from zero in D for all μ sufficiently small. For the equation (1.6) this condition reduces to $\frac{\partial f}{\partial y} \leq -\bar{L} < 0$ where $\bar{L}$ is a constant.

We will naturally assume that all of the conditions of the hypothesis of Theorem 1.2 are satisfied.

THEOREM 4.1. Let the consistent finite difference equation

$$(4.1) \quad \sum_{i=0}^k \alpha_i Y_{n+i} = h^{1-\gamma} \sum_{i=0}^k \beta_i F_{n+i} + R_n$$

where $F_{n+1} = f(Y_{n+1}, x_{n+1})$ and R_n is the round-off error satisfying the following conditions:

- a) $R_n = O(h^{2(1-\gamma)});$
- b) - The finite difference equation is relatively stable;
- c) $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$ and $\frac{\partial^2 f}{\partial^2 y}$ are continuous and bounded
 $(-\bar{L} \leq \frac{\partial f}{\partial y} \leq -\bar{L} < 0)$ for $0 \leq x \leq 1$ and $-\infty < y < \infty;$
- d) $|e_i| = |Y_i - y_i| \leq Th^{1-\gamma}$ for $i = 0, 1, \dots, k-1$ where T is a positive constant independent of h .

Then there exists a constant C for a given value of h such that $|e_n| \leq Ch^{1-\gamma}$ for $n = 0, 1, \dots, N$ where $0 \leq x_n \leq x_N = 1$.

Proof: The exact solution of $\mu y' = f(y, x)$ satisfies the difference equation

$$(4.2) \quad \sum_{i=0}^k \alpha_i y_{n+i} = h^{1-\gamma} \sum_{i=0}^k \beta_i f_{n+i} + T_n$$

where T_n is the truncation error and $f_{n+i} = f(y(x_{n+i}), x_{n+i})$. By our hypothesis on $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ and the consistency condition we can show that $T_n = O(h^{2-2\gamma})$ as $h \rightarrow 0$. This is done as follows: By Taylor series we have

$$y_{n+i} = y_n + ih y_n' + \frac{(ih)^2}{2!} y''(\theta_{n+i})$$

where

$$x_n \leq \theta_{n+i} \leq x_{n+i}$$

and

$$\frac{f_{n+i}}{h^\gamma} = y'_{n+i} = y'_n + ihy''(\theta'_{n+i})$$

where

$$x_n \leq \theta'_{n+i} \leq x_{n+i}.$$

Also,

$$y'' = \frac{d}{dx} \left(\frac{f(y, x)}{\mu} \right) = f_y \frac{y'}{\mu} + \frac{f_x}{\mu} = \frac{f_y f}{\mu^2} + \frac{f_x}{\mu}.$$

Substituting these expressions into (4.2), we obtain

$$\begin{aligned} y_n \sum_{i=0}^k \alpha_i + hy'_n \sum_{i=0}^k (i\alpha_i - \beta_i) \\ + \sum_{i=0}^k \alpha_i \frac{(ih)^2}{2} y''(\theta'_{n+i}) - h^2 \sum_{i=0}^k \beta_i (iy''(\theta'_{n+i})) \\ = T_n. \end{aligned}$$

Or since $\sum_{i=0}^k \alpha_i = 0$ and $\sum_{i=0}^k (i\alpha_i - \beta_i) = 0$ by the consistency condition, we have

$$\begin{aligned} T_n = h^2 \sum_{i=0}^k \frac{\alpha_i}{2} (i)^2 \left[\frac{f_y(y(\theta'_{n+i}), \theta'_{n+i})}{h^{2\gamma}} f(y(\theta'_{n+i}), \theta'_{n+i}) \right. \\ \left. + \frac{f_x(y(\theta'_{n+i}), \theta'_{n+i})}{h^\gamma} \right] \end{aligned}$$

$$\begin{aligned}
& - h^2 \sum_{i=0}^k i \beta_i \left[\frac{f_y(y(\theta'_{n+i}), \theta'_{n+i})}{h^{2\gamma}} f(y(\theta'_{n+i}), \theta'_{n+i}) \right. \\
& \quad \left. + \frac{f_x(y(\theta'_{n+i}), \theta'_{n+i})}{h^\gamma} \right] \\
& = O(h^{2-2\gamma}) \quad \text{as } h \rightarrow 0.
\end{aligned}$$

We now return to the proof of the theorem. Subtracting equation (4.2) from (4.1) and letting $e_i = Y_i - y_i$, we obtain

$$\begin{aligned}
(4.3) \quad \sum_{i=0}^k \alpha_i e_{n+i} &= h^{1-\gamma} \sum_{i=0}^k \beta_i (F_{n+i} - f_{n+i}) + R_n - T_n \\
&= h^{1-\gamma} \sum_{i=0}^k \beta_i \frac{\partial \bar{f}_{n+i}}{\partial y} e_{n+i} + R_n - T_n
\end{aligned}$$

where $\frac{\partial \bar{f}_{n+i}}{\partial y} = \frac{\partial f}{\partial y}(y_{n+i} + \delta_{n+i}(Y_{n+i} - y_{n+i}), x_{n+i})$, i.e.,

$\frac{\partial f}{\partial y}$ is evaluated at a point between y_{n+i} and Y_{n+i} .

If we denote the right side of (4.3) by Q_n , we may proceed as in Theorem 2.2 using the following bound for Q_n

$$\begin{aligned}
(4.4) \quad |Q_n| &\leq h^{1-\gamma} \sum_{i=0}^k |\beta_i| \left| \frac{\partial \bar{f}_{n+i}}{\partial y} \right| |e_{n+i}| + |R_n - T_n| \\
&\leq h^{1-\gamma} \bar{L}_\beta \sum_{i=0}^k |e_{n+i}| + |R_n - T_n|
\end{aligned}$$

where

$$\beta = \max_{i=0,1,\dots,k} |\beta_i|.$$

It will be recalled that in Theorem 2.2 for $0 \leq n \leq N_0 = \frac{\ell n h^{1-\gamma}}{\ell n r}$ where $r = \frac{1}{2}(1 + \max_{|\zeta_{j0}| < 1} |\zeta_{j0}|)$, it was

possible to obtain an explicit bound for e_n . If we replace the L_0 in Theorem 2.2 by $\bar{L}$, we obtain the following bound for e_n in the non-linear case

$$\begin{aligned}
 |e_n| &\leq K_3 h^{1-\gamma} (1 + Ah^{1-\gamma})^n \\
 &\leq K_3 h^{1-\gamma} e^{nAh^{1-\gamma}} \\
 (4.5) \quad |e_n| &\leq K_3 h^{1-\gamma} e^{Ah^{1-\gamma} \frac{\ell n h^{1-\gamma}}{\ell n r}} \quad \text{for } 0 \leq n \leq N_0
 \end{aligned}$$

$$\text{where } K_3 = \frac{gK_2 N_0 h^{1-\gamma} + K_1}{1 - h_0^{1-\gamma} \beta g \bar{L}} \quad \text{and} \quad A = \frac{\beta g \bar{L} (k+1)}{1 - h_0^{1-\gamma} \beta g \bar{L}}$$

where $h_0^{1-\gamma} \beta g \bar{L} < 1$ and $h \leq h_0$. K_2 is a bound for R_n and T_n ; g is the Green's function bound, and $K_1 = k^2 \alpha g T$, the bound due to the initial error, where $\alpha = \max |\alpha_i|$.

We now find an interval $[0, x_{N_1}]$ where $-\bar{L}_1 \leq \frac{\partial f(y(x), x)}{\partial y} \leq -\bar{L}_1$ and

$$(4.6) \quad h^{1-\gamma} \sum_{i=0}^{k-1} (\bar{L}_1 - L_1) \sum_{j=0}^k |\beta_j| |c_i| \frac{(1 - |\zeta_i|^{N_1})}{1 - |\zeta_i|} < 1$$

where the $\sum$ means the sum is taken over those i 's corresponding to roots $|\zeta_{i0}| = 1$ and the c_i 's are defined with respect to the difference equation

$$\sum_{i=0}^k (\alpha_i + h^{1-\gamma} \beta_i \bar{L}_1) \Phi_{n+i} = 0$$

with initial conditions $\Phi_0 = \Phi_1 = \dots = \Phi_{k-2} = 0$ and $\Phi_{k-1} = \frac{1}{\alpha_k + h^{1-\gamma} \beta_k \bar{L}_1}$. The Green's function takes the form

$$\Phi_n = c_0 \zeta_0^n + c_1 \zeta_1^n + \dots + c_{k-1} \zeta_{k-1}^n$$

and by relative stability

$$|\zeta_i| \leq |\zeta_0| \quad i = 1, 2, \dots, k-1$$

where

$$\zeta_0 = 1 - \bar{L}_1 h^{1-\gamma} + \mathcal{O}(h^{2(1-\gamma)}).$$

By a close examination of (4.6) it becomes clear that the length of the interval $[0, x_{N_1}]$ depends upon the value of

$$(4.7) \quad 1 \leq \sum_{i=0}^{k-1} |c_i| \sum_{j=0}^k |\beta_j|.$$

We shall see later how to choose the α_i and β_i to make this sum close to 1. For now we will merely assume that the sum is close enough to one so that $N_0 < N_1$.

If we now add $h^{1-\gamma} \sum_{i=0}^k \beta_i \bar{L}_1 e_{n+i}$ to both sides of equation (4.3), we obtain

$$(4.8) \quad \sum_{i=0}^k (\alpha_i + h^{1-\gamma} \beta_i \bar{\bar{L}}_1) e_{n+i} =$$

$$h^{1-\gamma} \sum_{i=0}^k \beta_i (\bar{\bar{L}}_1 + \frac{\partial \bar{f}}{\partial y} e_{n+i}) e_{n+i} + R_n - T_n.$$

We let the right hand side of (4.8) be denoted by q_n , and we note that

$$|\bar{\bar{L}}_1 + \frac{\partial f(y_{n+i} + \delta_{n+i} e_{n+i}, x_{n+i})}{\partial y}|$$

$$\leq |\bar{\bar{L}}_1 + \frac{\partial f(y_{n+i}, x_{n+i})}{\partial y} + \frac{\partial^2 \bar{f}}{\partial y^2} \delta_{n+i} e_{n+i}|$$

$$\leq \bar{\bar{L}}_1 - \bar{L}_1 + \mathcal{O}(e_{n+i}) \quad \begin{array}{l} \text{for } 0 \leq n \leq N_1 - k \\ \text{and } i = 0, 1, \dots, k \end{array}$$

where we have used condition c), i.e., $\frac{\partial^2 \bar{f}}{\partial y^2}$ is continuous and bounded.

Now it follows that

$$(4.9) \quad |q_i| \leq h^{1-\gamma} (\bar{\bar{L}}_1 - \bar{L}_1) \sum_{j=0}^k |\beta_j| |e_{i+j}| +$$

$$+ \mathcal{O}(h^{2(1-\gamma)}) + \mathcal{O}(h^{1-\gamma} \sum_{j=0}^k |e_{i+j}|^2)$$

$$0 \leq i \leq N_1 - k$$

using condition a) and our knowledge about the behavior of the truncation error.

The solution to the difference equation (4.8) is

$$(4.10) \quad e_n = \begin{cases} \sum_{i=0}^{n-k} \bar{\Phi}_{n-i-1} q_i + \Psi_n & n \geq k \\ \Psi_n & n < k \end{cases}$$

where $\bar{\Phi}_n$ satisfies the difference equation

$$\sum_{i=0}^k (\alpha_i + h^{1-\gamma} \beta_i \bar{\bar{L}}_1) \bar{\Phi}_{n+i} = 0$$

with initial conditions $\bar{\Phi}_0 = \bar{\Phi}_1 = \dots = \bar{\Phi}_{k-2} = 0$ and $\bar{\Phi}_{k-1} = (\alpha_k + h^{1-\gamma} \beta_k \bar{\bar{L}}_1)^{-1}$.

Ψ_n is given by

$$\Psi_n = \sum_{i=0}^{k-1} \left(\sum_{j=0}^{k-i-1} (\alpha_{k-j} + h^{1-\gamma} \beta_{k-j} \bar{\bar{L}}_1) \bar{\Phi}_{n+k-i-j-1} \right) e_i$$

for $n \geq k$.

However, for $n \geq N_0$ we have by Lemma 2 that

$$|\bar{\Phi}_n| \leq \bar{\Phi} \left(\frac{r^{N_0}}{h^{(1-\gamma)(m-1)/m}} + (1 - Lh^{1-\gamma})^{N_0} \right) \quad N_0 \leq n$$

where L , r , and $\bar{\Phi}$ are uniform bounds for all $\bar{\bar{L}}_1$ between $\bar{\bar{L}}$ and $\bar{L}$, and where m equals the maximum multiplicity of ζ_{j_0} . Therefore, using condition d)

$$\begin{aligned}
|\Psi_n| &\leq (\alpha + h^{1-\gamma} \beta \bar{\bar{L}}) \times (|\Phi_{n+k-1}| + \dots + |\Phi_n|) \sum_{i=0}^{k-1} |e_i| \\
&\leq k^2 (\alpha + h^{1-\gamma} \beta \bar{\bar{L}}) \Phi \left(\frac{r^{N_0}}{h^{1-\gamma}} + (1 - Lh^{1-\gamma})^{N_0} \right) Th^{1-\gamma}
\end{aligned}$$

$$(4.11) \quad |\Psi_n| \leq K_4 Th^{1-\gamma} \quad \text{for } n \geq N_0$$

where

$$K_4 \geq k^2 (\alpha + h^{1-\gamma} \beta \bar{\bar{L}}) \Phi \left(\frac{r^{N_0}}{h^{1-\gamma}} + (1 - Lh^{1-\gamma})^{N_0} \right).$$

By using Lemma 1 we can find a bound for the part of (4.10) that is due to the round-off and truncation errors:

$$(4.12) \quad \sum_{i=0}^{n-k} |\Phi_{n-i-1}| \mathcal{O}(h^{2(1-\gamma)}) \leq K_5 h^{1-\gamma}$$

where K_5 can be chosen as a uniform bound for all $\bar{\bar{L}}_1$ such that $-\bar{\bar{L}} \leq \bar{\bar{L}}_1 \leq -\bar{\bar{L}} \leq 0$.

If we define $E_n = \max(|e_0|, |e_1|, \dots, |e_n|)$, then by using Lemma 1 again

$$\sum_{i=0}^{n-k} |\Phi_{n-i-1}| \mathcal{O}(h^{1-\gamma} E_n^2) \leq K_6 E_n^2.$$

A bound can now be found for e_n using the last inequality, (4.9), (4.11), and (4.12) in equation (4.10):

$$\begin{aligned}
|e_n| &\leq \sum_{i=0}^{n-k} |\Phi_{n-i-1}| |q_i| + |\Psi_n| \\
&\leq h^{1-\gamma} (\bar{\bar{L}}_1 - \bar{L}_1) \sum_{j=0}^k |\beta_j| E_n \sum_{i=0}^{n-k} |\Phi_{n-i-1}| \\
&\quad + K_4 Th^{1-\gamma} + K_5 h^{1-\gamma} + K_6 E_n^2 \\
&\leq h^{1-\gamma} (\bar{\bar{L}}_1 - \bar{L}_1) \sum_{j=0}^k |\beta_j| E_n \sum_{i=0}^{n-k} (|c_0| |\zeta_0|^{n-i-1} + \\
&\quad \dots + |c_{k-1}| |\zeta_{k-1}|^{n-i-1}) + K_4 Th^{1-\gamma} + K_5 h^{1-\gamma} + \\
&\quad K_6 E_n^2 \\
&\leq \left[\sum_{i=0}^{k-1} h^{1-\gamma} (\bar{\bar{L}}_1 - \bar{L}_1) \sum_{j=0}^k |\beta_j| |c_i| \frac{(1-|\zeta_i|^n)}{1-|\zeta_i|} + \right. \\
&\quad \left. O(h^{\frac{1-\gamma}{m}}) \right] E_n + K_4 Th^{1-\gamma} + K_5 h^{1-\gamma} + K_6 E_n^2
\end{aligned}$$

where the $O(h^{(1-\gamma)/m})$ term results from the roots $|\zeta_{j0}| < 1$ and Lemma 2, i.e.

$$|\Phi_n| \leq \Phi \left(\frac{r^n}{h^{(1-\gamma)(m-1)/m}} + (1 - Lh^{1-\gamma})^n \right)$$

where $r < 1$.

By our choice of $\bar{\bar{L}}_1$ and $\bar{L}_1$ (equation (4.6)) and assuming h is sufficiently small the term in brackets

multiplying E_n will be less than $a < 1$.

Hence,

$$(4.13) \quad |e_n| \leq aE_n + K_5 h^{1-\gamma} + K_4 T h^{1-\gamma} + K_6 E_n^2 \quad \text{for } N_0 \leq n \leq N_1.$$

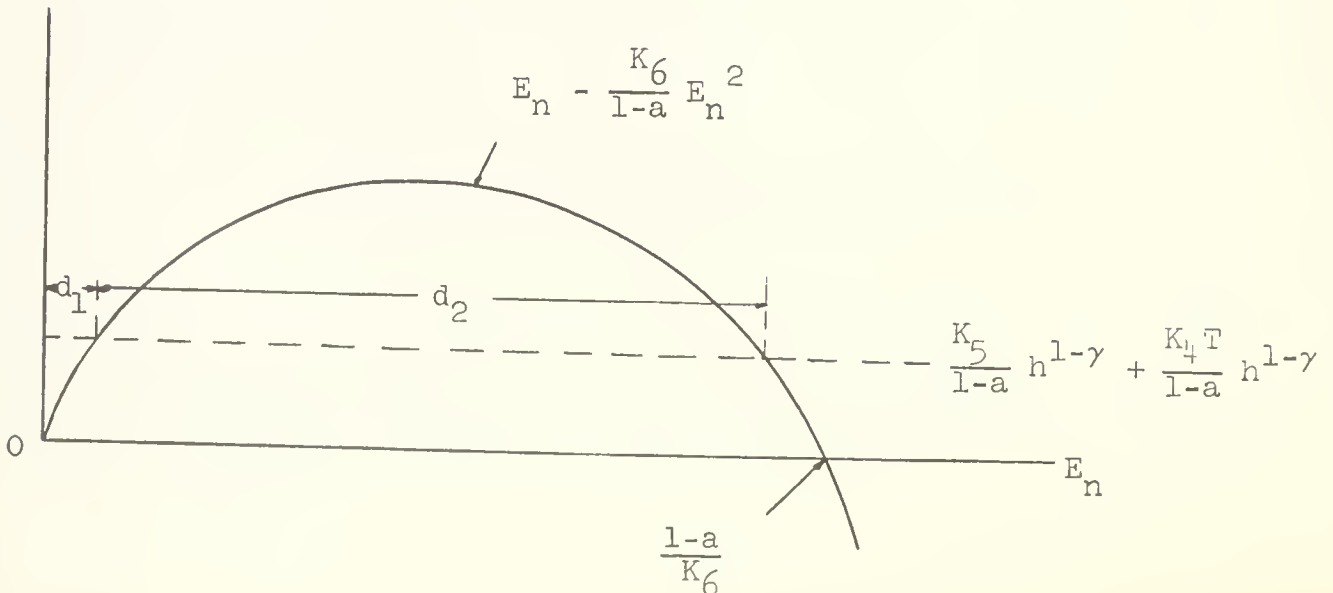
If $E_n = |e_j|$ for $j \leq N_0$, then we may use our estimate (4.5). If $j > N_0$, we can replace $|e_n|$ in (4.13) by E_n and obtain

$$(4.14') \quad E_n \leq \frac{K_5}{1-a} h^{1-\gamma} + \frac{K_4 T h^{1-\gamma}}{1-a} + \frac{K_6 E_n^2}{1-a}$$

for $0 \leq n \leq N_1$

since this bound is larger than the one for $|e_n|$ for $0 \leq n \leq N_0$, i.e., see (4.5).

Geometrically, we have the following picture:



We wish to show that $E_n \leq d_1$ for $0 \leq n \leq N_1$. It will then follow that

$$d_1 - \frac{K_6}{1-a} d_1^2 - \frac{K_5 h^{1-\gamma}}{1-a} - \frac{K_4 T h^{1-\gamma}}{1-a} = 0$$

or

$$\begin{aligned} d_1 &= \frac{-1 + \sqrt{1 - 4 \frac{(K_4 T + K_5)}{(1-a)^2} K_6 h^{1-\gamma}}}{\frac{-2K_6}{1-a}} \\ &= \frac{(K_4 T + K_5) h^{1-\gamma}}{1-a} + \mathcal{O}(h^{2(1-\gamma)}). \end{aligned}$$

The proof that $E_n \leq d_1$ for $0 \leq n \leq N_1$ is by induction. Rewriting the difference equation (4.3) with n replaced by $n-k$, we have

$$\sum_{i=0}^k \alpha_i e_{n-k+i} = h^{1-\gamma} \sum_{i=0}^k \beta_i \frac{\partial \bar{f}}{\partial y}{}^{n-k+i} e_{n-k+i} + R_{n-k} - T_{n-k}.$$

For $h \leq h_0$ where $|\alpha_k| > \bar{L} |\beta_k| h_0^{1-\gamma}$ we have

$$\begin{aligned} |e_n| &\leq \frac{\sum_{i=0}^{k-1} |-\alpha_i + h^{1-\gamma} \beta_i \frac{\partial \bar{f}}{\partial y}{}^{n-k+i}| |e_{n-k+i}|}{|\alpha_k| - \bar{L} |\beta_k| h_0^{1-\gamma}} \\ &\quad + \frac{|R_{n-k} - T_{n-k}|}{|\alpha_k| - \bar{L} |\beta_k| h_0^{1-\gamma}} \end{aligned}$$

$$\leq K_7 E_{n-1} + K_8 h^{2(1-\gamma)}$$

where

$$\sum_{i=0}^{k-1} \frac{|\alpha_i + h^{1-\gamma} \beta_i \frac{\partial \bar{f}}{\partial y}|}{|\alpha_k| - \bar{L} |\beta_k| h_0^{1-\gamma}} \leq K_7$$

and

$$\frac{|R_{n-k} - T_{n-k}|}{|\alpha_k| - \bar{L} |\beta_k| h_0^{1-\gamma}} \leq K_8 h^{2(1-\gamma)}.$$

$E_{k-1} \leq Th^{1-\gamma} < d_1$ by hypothesis. By induction assume that $E_{n-1} \leq d_1$ and h is so small that

$$|e_n| \leq K_7 d_1 + K_8 h^{2(1-\gamma)} < d_1 + d_2$$

and therefore $E_n < d_1 + d_2$. But now by (4.14') we have that $E_n < d_1 + d_2$ implies $E_n \leq d_1$. Therefore, by induction

$$(4.14) \quad E_n \leq \frac{K_5 h^{1-\gamma}}{1-a} + \frac{K_4 Th^{1-\gamma}}{1-a} + O(h^{2(1-\gamma)}) \quad \text{for } 0 \leq n \leq N_1.$$

We may now repeat the argument and extend the interval to $[0, x_{N_2}]$ where $N_1 < N_2$. We find an interval $[x_{N_1-k+1}, x_{N_2}]$ where $-\bar{L}_2 \leq \frac{\partial f}{\partial y} \leq -\bar{L}_2$ and

$$(4.15) \quad \sum_{i=0}^{k-1} h^{1-\gamma} (\bar{\bar{L}}_2 - \bar{L}_2) \sum_{j=0}^k |\beta_j| |c_i| |\zeta_i|^{k-1} \frac{(1-|\zeta_i|^{N_2+1-N_1})}{1-|\zeta_i|}$$

$$< 1$$

where the C_i 's are now defined with respect to the Green's function associated with the difference equation

$$\sum_{i=0}^k (\alpha_i + h^{1-\gamma} \beta_i \bar{\bar{L}}_2) \Phi_{n+i} = 0$$

with initial conditions $\Phi_0 = \Phi_1 = \dots = \Phi_{k-2} = 0$ and

$$\Phi_{k-1} = (\alpha_k + h^{1-\gamma} \beta_k \bar{\bar{L}}_2)^{-1}.$$

We add $h^{1-\gamma} \sum_{i=0}^k \beta_i \bar{\bar{L}}_2 e_{n+i}$ to both sides of equation (4.3):

$$\sum_{i=0}^k (\alpha_i + h^{1-\gamma} \beta_i \bar{\bar{L}}_2) e_{n+i} = h^{1-\gamma} \sum_{i=0}^k \beta_i \left(\bar{\bar{L}}_2 + \frac{\partial \bar{f}}{\partial y}^{n+i} \right) e_{n+i} + R_n - T_n.$$

We treat this equation exactly as (4.8) was treated but we use (4.14) for the initial estimate. The accumulative error for the first two intervals is

$$E_n \leq \frac{K_5}{1-a} h^{1-\gamma} + \frac{K_4}{1-a} \left[\frac{K_5}{1-a} h^{1-\gamma} + \frac{K_4}{1-a} T h^{1-\gamma} \right] + O(h^{2(1-\gamma)})$$

$$\begin{aligned}
&\leq \left[1 + \frac{K_4}{1-a}\right] \frac{K_5}{1-a} h^{1-\gamma} + \left(\frac{K_4}{1-a}\right)^2 Th^{1-\gamma} + O(h^{2(1-\gamma)}) \\
&= \frac{\left[1 - \left(\frac{K_4}{1-a}\right)^2\right]}{\left[1 - \frac{K_4}{1-a}\right]} \frac{K_5}{1-a} h^{1-\gamma} + \left(\frac{K_4}{1-a}\right)^2 Th^{1-\gamma} + O(h^{2(1-\gamma)})
\end{aligned}$$

$$0 \leq n \leq N_2.$$

At the end of the third interval the error will be

$$\begin{aligned}
E_n &\leq \frac{K_5}{1-a} h^{1-\gamma} + \frac{K_4}{1-a} \frac{\left[1 - \left(\frac{K_4}{1-a}\right)^2\right]}{1 - \frac{K_4}{1-a}} \frac{K_5}{1-a} h^{1-\gamma} \\
&\quad + \left(\frac{K_4}{1-a}\right)^3 Th^{1-\gamma} + O(h^{2(1-\gamma)}) \\
&= \frac{\left[1 - \left(\frac{K_4}{1-a}\right)^3\right]}{1 - \frac{K_4}{1-a}} \frac{K_5}{1-a} h^{1-\gamma} + \left(\frac{K_4}{1-a}\right)^3 Th^{1-\gamma} + O(h^{2(1-\gamma)}) \\
&\quad 0 \leq n \leq N_3.
\end{aligned}$$

Finally at the end of the ℓ interval we obtain

$$(4.16) \quad E_n \leq \frac{\left[1 - \left(\frac{K_4}{1-a}\right)^\ell\right]}{\left[1 - \frac{K_4}{1-a}\right]} \frac{K_5}{1-a} h^{1-\gamma} + \left(\frac{K_4}{1-a}\right)^\ell Th^{1-\gamma} + O(h^{2(1-\gamma)})$$

for $0 \leq n \leq N_\ell = N$. Q.E.D.

Of course, the bound for E_n given by (4.16) may be large if l is much greater than 1, i.e. if

$$\sum_{i=0}^{k-1} |c_i| \sum_{j=0}^k |\beta_j| \text{ is much greater than } 1.$$

We therefore wish to minimize this sum. $\sum_{j=0}^k |\beta_j|$ will be a minimum if we choose all $\beta_j \geq 0$. Since the roots $\zeta_{j0} \neq 1$ on the unit circle do not yield μ -stable difference schemes of higher precision than those roots inside the unit circle, we will exclude such roots for now.

We are thus left with the problem of minimizing $|c_0| \sum_{j=0}^k |\beta_j|$. The following corollaries will be very useful:

COROLLARY 1. If the only essential root is $\zeta_{j0} = 1$, and if $\alpha_k = 1$ and $\beta_j \geq 0$ for $j = 0, 1, \dots, k$, then the value of l in Theorem 4.1 is one.

Proof: Since $\beta_j \geq 0$ by consistency

$$\begin{aligned} \rho'(1) &= \sum_{j=0}^k |\beta_j| \\ &= (1 - r_1)(1 - r_2) \cdots (1 - r_{k-1}) \end{aligned}$$

where

$$\rho(r_i) = 0 \quad i = 1, 2, \dots, k-1.$$

In Lemma 1 it was shown that

$$|c_0| \leq \frac{\frac{1}{\alpha_k + \mathcal{O}(h^{1-\gamma})}}{|(\xi_0 - \xi_1)(\xi_0 - \xi_2) \cdots (\xi_0 - \xi_{k-1})|}$$

$$\approx \frac{1 + \mathcal{O}(h^{1-\gamma})}{\rho'(1)}.$$

Therefore,

$$|c_0| \sum_{j=0}^k |\beta_j| \approx 1$$

and ℓ can be chosen equal to 1. Q.E.D.

COROLLARY 2. If the only essential root is $\zeta_{j0} = 1$ and

$$\frac{(\bar{\bar{L}} - \bar{L})|c_0|}{\bar{\bar{L}} + \mathcal{O}(h^{1-\gamma})} \sum_{j=0}^k |\beta_j| < 1$$

where

$$|\zeta_0| = 1 - \bar{\bar{L}}h^{1-\gamma} + \mathcal{O}(h^{2(1-\gamma)}),$$

then the value of ℓ in Theorem 4.1 is one.

Proof: The proof is obvious since (4.6) is satisfied for all $0 \leq x_n \leq 1$.

We will consider a few examples to illustrate how the value of l may be estimated in practice. Consider the differential equation

$$\mu y' = -y(y-1)(20x+10); \quad y(0) = 2.$$

In the boundary layer for all $0 < h \leq h_0$

$$-30 \leq f_y \leq -10.$$

Example 1. Suppose Adam's method

$$Y_{n+3} - Y_{n+2} = \frac{h^{1-\gamma}}{24} (9F_{n+3} + 19F_{n+2} - 5F_{n+1} + F_n)$$

is used to calculate the solution to the above differential equation. Here

$$|c_0| \approx \sum |\beta_i| = \frac{34}{24}.$$

Since f_y is monotonic in the boundary layer $\bar{L}_1 = \bar{\bar{L}}_2 = L_1$, $\bar{L}_2 = \bar{\bar{L}}_3 = L_2$, etc. $\frac{(30 - L_1)}{30} \frac{34}{24} = .8$ where we have let $a = .8 < 1$.

$$\begin{aligned} L_1 &= 30 \left(1 - .8 \left(\frac{24}{34} \right) \right) \\ &= 30 (.435) = 13.1. \end{aligned}$$

$$L_2 = 30 (.435)^2 = 5.66$$

$$L_\ell = 30 (.435)^\ell.$$

Since $L_2 = 5.66$, it is clear that $\ell = 2$ is a good estimate for the value of ℓ in the boundary layer.

Example 2. Consider all methods with essential root $\zeta = 1$ and

$$|c_0| \leq 2.$$

We choose $a = .8$ again. Then

$$L_\ell = 30 (.6)^\ell$$

$$L_1 = 18$$

$$L_3 = 6.5 .$$

Thus we take $\ell = 3$ in the boundary layer.

Example 3. The same as Example 2, but let $|c_0| \leq 3$.

Here

$$L_\ell = 30 (.734)^\ell$$

$$L_4 = 8.7 .$$

Therefore, we choose $\ell = 4$.

The above calculations were made inside of the boundary layer; outside of the boundary layer we make use of the following fact from the asymptotic theory:

$$f_y(y(x,\mu),x) = f_y(\phi(x),x) + \mathcal{O}(h^\gamma)$$

for

$$-\frac{2}{\bar{L}} h^\gamma \ln h^\gamma \leq x \leq 1.$$

In our examples $\phi(x) = 1$ so

$$f_y(1,x) = -20x - 10$$

outside the boundary layer. Here $f_y(1,x)$ is independent of h and is monotonically decreasing from -10 to -30 . Therefore, in Examples 1, 2, and 3 we must increase ℓ by 2, 3, and 4 respectively in order to include the entire interval $[0,1]$ in our estimate.

In practice it was observed that there was no error build up outside of the boundary layer region for μ -stable schemes. Therefore, the above estimates are to be considered as absolute maxima for the value of ℓ .

A few remarks about Theorem 4.1 are in order:

Remark 1: The same proof could be used to obtain bounds for μ -stable methods instead of the less general relatively stable methods, but these μ -stable methods would require a much larger value of ℓ .

Remark 2: Condition to (c) could be weakened by requiring that the functions $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$ and $\frac{\partial^2 f}{\partial y^2}$ be continuous in a strip: $0 \leq x \leq 1$ and $|y - y(x)| < t$ where t is as large as is necessary in the proof.

Remark 3: By using the variational equation and by weakening conditions (a), (c), and (d) slightly we can gain another $O(h^{1-\gamma})$ in our error bounds. This is done as follows: We make the following changes in the hypothesis of Theorem 4.1:

$$(a') \quad R_n = O(h^{3(1-\gamma)})$$

(c') $f(y, x)$ has two continuous and bounded derivatives for $0 \leq x \leq 1$ and $-\infty < y < \infty$.

$$(d') \quad e_i = O(h^{2(1-\gamma)}) \quad i = 0, 1, \dots, k-1$$

Using conditions (a'), (c') and (d') we derive the so-called variational equation.

Using condition (c') and (4.16), we may write

$$f(Y_n, x_n) - f(y(x_n), x_n) = f_y(y(x_n), x_n)e_n + O(h^{2(1-\gamma)}).$$

Our difference equation (4.3) takes the form

$$(4.17) \quad \sum_{i=0}^k \alpha_i e_{n+i} = h^{1-\gamma} \left[\sum_{i=0}^k \beta_i f_y(y(x_{n+i}), x_{n+i}) e_{n+i} + \right. \\ \left. O(h^{2(1-\gamma)}) \right] - Cy''(x_n)h^2 + O(h^{3(1-\gamma)})$$

where condition (a') was used for the round-off error and $Cy''(x_n)h^2$ for the truncation error.

If we let $\bar{e}_n = \frac{e_n}{h^{1-\gamma}}$, equation (4.17) becomes

$$(4.18) \quad \sum_{i=0}^k \alpha_i \bar{e}_{n+i} = h^{1-\gamma} \left[\sum_{i=0}^k \beta_i f_y(y(x_{n+i}), x_{n+i}) \bar{e}_{n+i} \right] \\ - Cy''(x_n)h^{1+\gamma} + O(h^{2(1-\gamma)}).$$

Since $h^{2\gamma}y''(x_n)$ is a continuous function for all μ , we can write

$$h^{2\gamma} Cy''(x_n) = C h^{2\gamma} \sum_{i=0}^k \beta_i y''(x_{n+i}) + O(h).$$

Substituting the last expression into (4.18) gives

$$(4.19) \quad \sum_{i=0}^k \alpha_i \bar{e}_{n+i} = h^{1-\gamma} \sum_{i=0}^k \beta_i (f_y(y(x_{n+i}), x_{n+i}) \bar{e}_{n+i} \\ - C h^{2\gamma} y''(x_{n+i})) + O(h^{2(1-\gamma)}).$$

This is the difference equation that would result if the initial value problem, the variational equation,

$$(4.20) \quad \mu e'(x) = f_y(y, x) e(x) - C \mu^2 y''(x),$$

$$e(0) = 0$$

were solved by the finite difference methods under discussion.

The initial conditions for $\bar{e}_i$ are $\bar{e}_i = O(h^{1-\gamma})$, $i = 0, 1, \dots, k-1$, where we have made use of (d'). These values differ from the corresponding values of $e(x_i)$ ($i = 0, 1, \dots, k-1$) by at most $O(h^{1-\gamma})$. Hence by the proof of Theorem 4.1 (or (4.16)) we can write

$$\bar{e}_n = e(x_n) + O(h^{1-\gamma}) \quad \text{for } 0 \leq n \leq N$$

$$(4.21) \quad e_n = e(x_n) h^{1-\gamma} + O(h^{2(1-\gamma)}) \quad \text{for } 0 \leq n \leq N.$$

In order to obtain an estimate for $e(x)$ we must first obtain an estimate for $\mu^2 y''(x)$. In Vasil'eva [4] it was shown that under the conditions of Theorem 1.2 and condition (c') $y(x, \mu)$ has an asymptotic expansion in μ of the form

$$y(x, \mu) \sim \phi(x) + \theta(x, \mu) + O(\mu)$$

where $\phi(x)$ is a smooth function and $|\theta(x, \mu)| < C e^{-Lx/\mu}$.

Therefore,

$$\begin{aligned}
\mu y' &= f(y, x) = f(\phi(x) + \theta(x, \mu), x) + \mathcal{O}(\mu) \\
&= f_y \theta(x, \mu) + \mathcal{O}(\mu)
\end{aligned}$$

or

$$|\mu y'| \leq C\mu + C e^{-\bar{L}x/\mu}.$$

Similarly,

$$\begin{aligned}
(4.22) \quad |\mu^2 y''| &= |f_y \mu y' + \mu f_x| \\
&\leq C\mu + C e^{-\bar{L}x/\mu}.
\end{aligned}$$

Integrating the variational equation, (4.20), yields

$$\begin{aligned}
|e(x)| &\leq \int_0^x \frac{e^{-\bar{L}(x-t)/\mu}}{\mu} |C\mu^2 y''(t)| dt \\
&\leq \int_0^x \frac{e^{-\bar{L}(x-t)/\mu}}{\mu} (C\mu + C e^{-\bar{L}x/\mu}) dt \\
&\leq C\mu(1 - e^{-\bar{L}x/\mu}) + \frac{Cx}{\mu} e^{-\bar{L}x/\mu}.
\end{aligned}$$

Thus

$$(4.23) \quad |e(x)| \leq C\mu(1 - e^{-\bar{L}x/\mu}) + \frac{Cx}{\mu} e^{-\bar{L}x/\mu}.$$

Substituting (4.23) into (4.21) gives

$$(4.24) \quad |e_n| \leq C[\mu(1 - e^{-\bar{L}x/\mu}) + \frac{x}{\mu} e^{-\bar{L}x/\mu}]h^{1-\gamma} + O(h^{2(1-\gamma)}).$$

Now (4.24) will be an improvement over (4.16) if ℓ is small, i.e., if $\ell = 1, 2, 3$, or 4.

Remark 4: If in Theorem 4.1 R_n and T_n are $O(h^{(p+1)(1-\gamma)})$ and $e_i = O(h^{p(1-\gamma)})$ for $i = 0, 1, \dots, k-1$, then the same proof will lead to the result that $e_n = O(h^{p(1-\gamma)})$ for $n = 0, 1, \dots, N$.

We are interested in investigating the "best method" when the only essential root is $\zeta = 1$. By the "best method" we will mean the one which allows both ℓ and T_n , the truncation error, to be a minimum simultaneously.

If the schemes with the smallest truncation errors also have the property that $\beta_i \geq 0$ $i = 0, 1, \dots, k$, then by Corollary 1 ℓ will have the value 1. This is precisely the case when $k = 2$. By using the methods outlined in Chapter 3 on optimal methods we find that the "best methods" for the roots $\zeta = 1$ and $\zeta = r$ where $|r| < 1$ take the form

$$Y_{n+2} - (1+r)Y_{n+1} + r Y_n =$$

$$\frac{h^{1-\gamma}}{12} [(5+r)F_{n+2} + (8-8r)F_{n+1} + (-5r-1)F_n] + T_n$$

where

$$T_n = \frac{(1+r)}{2^4} y^{(IV)} h^4 .$$

Now if $-1 < r \leq -\frac{1}{5}$ all β_i 's will be greater than zero. Therefore, we merely pick a value of r close to -1 in order to make T_n small. Experimentally, it was found by Hull and Newbery [5] that $r = -.97156$ gave the smallest error for a particular ordinary differential equation with the property that $-\bar{L} \leq \frac{\partial f}{\partial y} \leq -\bar{L} < 0$.

For the case $k = 3$ let the roots be $\zeta = 1, r_1$, and r_2 with $|r_1| < 1$ and $|r_2| < 1$. After a few pages of calculations we find that the optimal methods are characterized by

$$\sigma(\zeta) = \beta_3 \zeta^3 + \beta_2 \zeta^2 + \beta_1 \zeta + \beta_0$$

where

$$\beta_3 = \frac{1}{2^4} [9 + r_1 + r_2 + r_1 r_2]$$

$$\beta_2 = \frac{1}{2^4} [19 - 13(r_1 + r_2) - 5r_1 r_2]$$

$$\beta_1 = \frac{1}{2^4} [-5 - 13(r_1 + r_2) + 19r_1 r_2]$$

$$\beta_0 = \frac{1}{2^4} [1 + r_1 + r_2 + 9r_1 r_2] .$$

Clearly $\beta_3 \geq 0$ for all $|r_1| < 1$, $|r_2| < 1$. For β_2 we observe that

$$19 - 5r_1r_2 > 14.$$

We must require then that

$$14 - 13(r_1 + r_2) \geq 0.$$

For example, if r_1 and r_2 lie in the shaded region of Figure 1, then $\beta_2 \geq 0$.

In order to guarantee that $\beta_1 \geq 0$ we must require $\operatorname{Re} r_1 \leq 0$, $\operatorname{Re} r_2 \leq 0$ and

$$-13(r_1 + r_2) \geq 5$$

or

$$19r_1r_2 \geq 5.$$

These conditions will be satisfied if r_1 and r_2 lie in the shaded region in Figure 2.

For the case β_0 we notice that if both r_1 and r_2 lie in the unit circle in the left hand plane then $\beta_0 \geq 0$.

It is clear then that the condition that guarantees $\beta_1 \geq 0$ is the most restrictive. For example, the Adam's method ($r_1 = r_2 = 0$) does not satisfy this condition as can easily be seen:

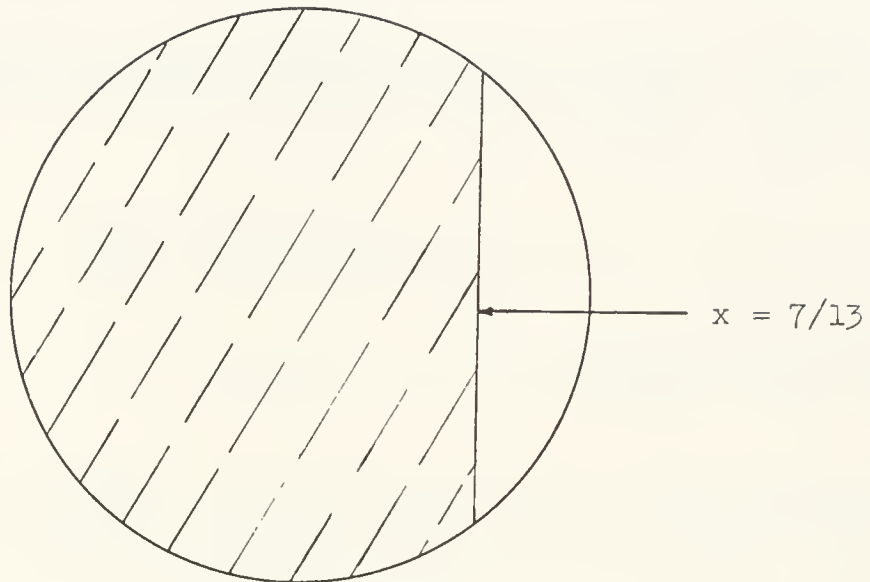


Figure 1

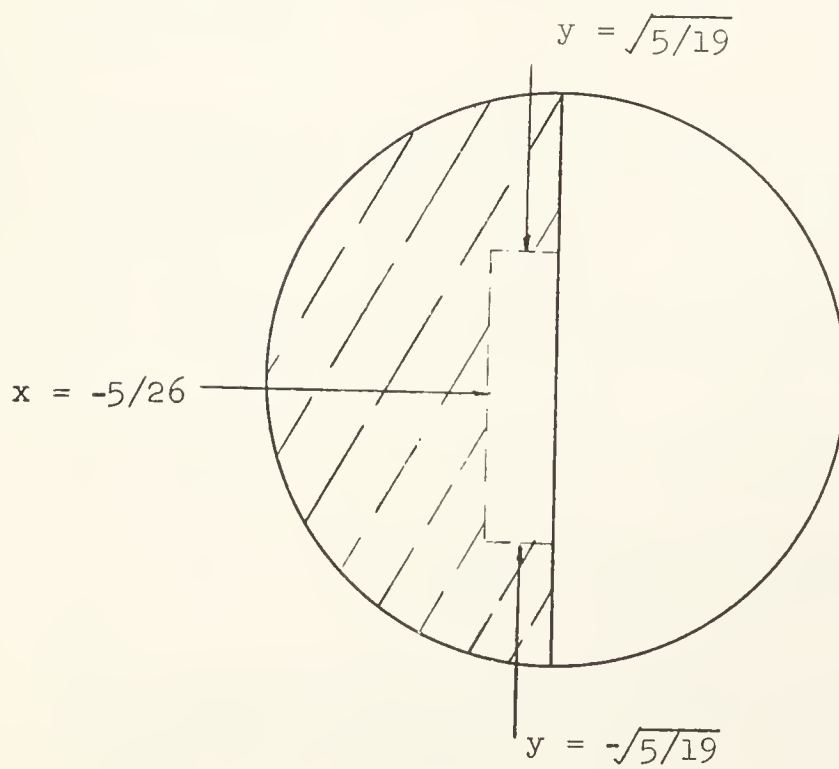


Figure 2

$$Y_{n+3} - Y_{n+2} = \frac{h^{1-\gamma}}{24} (9F_{n+3} + 19F_{n+2} - 5F_{n+1} + F_n).$$

We now investigate the truncation error for the case under study. Here

$$T_n = C_5 y^{(V)} h^5$$

where C_5 is given by

$$\begin{aligned} C_5 &= \frac{1}{5!} (\alpha_1 + 2^5 \alpha_2 + 3^5 \alpha_3) - \frac{1}{4!} (\beta_1 + 2^4 \beta_2 + 3^4 \beta_3) \\ &= \frac{1}{720} (-19 - 11r_1 - 11r_2 - 19r_1 r_2). \end{aligned}$$

The minimum occurs at

$$\frac{\partial C_5}{\partial r_1} = -19r_2 - 11 = 0 \quad \text{or} \quad r_2 = -\frac{11}{19}$$

$$\frac{\partial C_5}{\partial r_2} = -19r_1 - 11 = 0 \quad \text{or} \quad r_1 = -\frac{11}{19}$$

and

$$C_5 \Big|_{r_1=r_2=-\frac{11}{19}} = -\frac{1}{57}.$$

Fortunately, $r_1 = r_2 = -11/19$ lies in the region which guarantees that $\beta_i \geq 0$ $i = 0, 1, 2, 3$. If this were not the

case, we would have to accept a larger value of $|C_5|$.

For the case $k = 4$ we know from Dahlquist [1] and Henrici [2] that T_n is a minimum for the roots $\zeta = 1, -1, \zeta = e^{i\theta}$, and $\zeta = e^{-i\theta}$. This is clear since

$$T_n = \frac{1}{1440} [-27 - 27r_1r_2r_3 - 11(r_1 + r_2 + r_3) - 11(r_1r_2 + r_1r_3 + r_2r_3)] y^{(VI)}(x) h^6$$

where the roots are $\zeta = 1, r_1, r_2$, and r_3 . It therefore seems obvious that we should investigate the sign of β_i for the case $r_1 = -s$, $r_2 = se^{i\theta}$, and $r_3 = se^{-i\theta}$ where $0 < s < 1$. The corresponding β 's for the optimal methods are given by

$$\begin{aligned}\beta_4 &= \frac{1}{720} [251 + 19(r_1 + r_2 + r_3) + 11(r_1r_2 + r_1r_3 + r_2r_3) + 19r_1r_2r_3] \\ \beta_3 &= \frac{1}{720} [646 - 346(r_1 + r_2 + r_3) - 74(r_1r_2 + r_1r_3 + r_2r_3) - 106r_1r_2r_3] \\ \beta_2 &= \frac{1}{720} [-264 - 456(r_1 + r_2 + r_3) + 456(r_1r_2 + r_1r_3 + r_2r_3) + 264r_1r_2r_3] \\ \beta_1 &= \frac{1}{720} [106 + 74(r_1 + r_2 + r_3) + 346(r_1r_2 + r_1r_3 + r_2r_3) - 646r_1r_2r_3] \\ \beta_0 &= \frac{1}{720} [-19 - 11(r_1 + r_2 + r_3) - 19(r_1r_2 + r_1r_3 + r_2r_3) - 251r_1r_2r_3].\end{aligned}$$

Now $\beta_4 \geq 0$ for all r_1, r_2 , and r_3 with $|r_i| < 1$. On the other hand $\beta_3 \geq 0$ if $-1 < \operatorname{Re} r_i \leq 0$. This is true because

$$646 - 346(r_1 + r_2 + r_3) - 74(r_1 r_2 + r_1 r_3 + r_2 r_3) - 106r_1 r_2 r_3$$

$$\geq 646 - 74(3) - 106$$

$$> 0.$$

For the analysis of β_2 we set $r_1 = -s$, $r_2 = se^{i\theta}$, and $r_3 = se^{-i\theta}$ and $\frac{\pi}{2} \leq \theta \leq \pi$

$$-264 + 456s(1 - 2\cos\theta) + 456s^2(1 - 2\cos\theta) - 264s^3$$

$$\geq -528 + 456s + 456s^2.$$

Let

$$h(s) = -528 + 456s + 456s^2$$

$$h'(s) = 456 + 912s.$$

Thus for $0 \leq s \leq 1$, $h(s)$ is monotonically increasing. Also $h(.695) = 0$. Therefore, the allowable values of s are $.695 \leq s < 1$. Figure 3 shows the acceptable values for $r_1 = -s$, $r_2 = se^{i\theta}$, and $r_3 = se^{-i\theta}$.

$\beta_1 \geq 0$ for $-1 < \operatorname{Re} r_i \leq 0$ and all r_i chosen along the same radii because

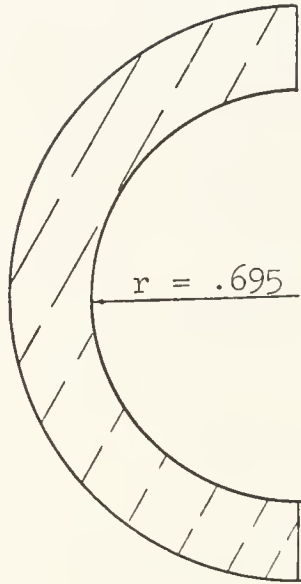


Figure 3

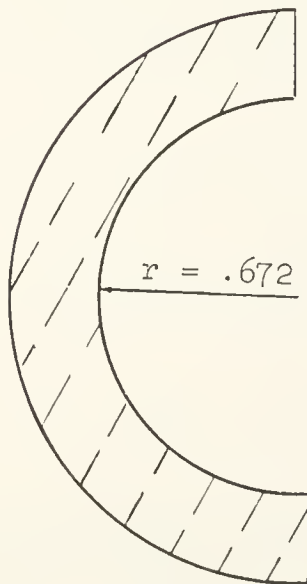


Figure 4

$$106 - 74s(1 - 2\cos\theta) + 346s^2(1 - 2\cos\theta) + 646s^3$$

$$\geq 106 - 222s + 346s^2 = h(s)$$

$$h'(s) = -222 + 692s = 0.$$

$h(s)$ is monotonically increasing if

$$s \geq \frac{222}{692} \approx .307.$$

But $h(.307) \geq 0$. For $0 \leq s \leq .307$ $h(s)$ is clearly greater than zero. Therefore, $h(s) \geq 0$ for $0 \leq s < 1$.

Finally, for β_0 our analysis is similar to the above where

$$-19 + 11s(1 - 2\cos\theta) - 19s^2(1 - 2\cos\theta) + 251s^3$$

$$\geq -19 + 11s - 57s^2 + 251s^3$$

$$\geq -76 + 11s + 251s^3 = h(s)$$

$$h'(s) = 11 + 753s^2.$$

Now if $s \geq \sqrt[3]{\frac{76}{251}} = .672$, $h(s) \geq 0$. This example is illustrated in Figure 4.

We therefore see that for the case $k = 4$ the requirement $\beta_2 \geq 0$ and $\beta_0 \geq 0$ (Figures 3 and 4 respectively) provide the limiting choice of r_1, r_2 , and r_3 .

As k increases the analysis becomes much more difficult and even calculating general expressions for the β_i 's and T_n in terms of $r_1, r_2, \dots, r_{k-1}$ is very tedious. We therefore resort to a slightly different approach.

From our analysis of $k = 2, 3$, and 4 we suspect that for the r_i 's in the negative half plane and near the unit circle there is some hope that for $k > 4$ all β_i 's will be greater than zero. We make use of the following formulas derived by Hull and Newbery [6] for optimal methods:

$$T_n = R \frac{y^{(k+2)} h^{k+2}}{(k+1)!} + O(h^{k+3})$$

$$R = \sum_{i=1}^k \bar{\alpha}_{i-1} \int_{i-1}^i x(x-1) \cdots (x-k) dx$$

where $\bar{\alpha}_{i-1} = \alpha_i + \alpha_{i+1} + \cdots + \alpha_k$ and

$$\beta_j = \sum_{i=1}^k \bar{\alpha}_{i-1} \int_{i-1}^i \frac{x(x-1) \cdots (x-k)}{(x-j)} dx$$

for $j = 0, 1, \dots, k$.

The above integrals can be calculated exactly by using the Newton-Cotes formulas. We have programmed the CDC6600

computer to calculate the values of β_j for the limiting values of r_i where we suspect favorable results for the β_j 's; that is, for k even let one root be at $\zeta = 1$, another at $\zeta = -1$, and all the remaining ones at $\zeta = \pm i$.

For k odd let one root be at $\zeta = 1$ and all the others at $\zeta = \pm i$. We refer to this choice of the roots as α -min. This is in contrast to α -max where one root is chosen at $\zeta = +1$ and the remaining ones at $\zeta = -1$.

Although both α -min and α -max define unstable schemes, in practice we would choose root at $\zeta = 1$ and the other roots inside the unit circle but near the roots of α -min or α -max when they lead to β_i 's ≥ 0 .

The results of the computer calculations are given in Table 1 for $k = 3, 4, \dots, 18$. By $\beta \geq 0$ we mean that $\beta_i \geq 0$ for $i = 0, 1, \dots, k$. $\beta < 0$ means that at least one β_i was less than zero.

We will conclude this section by saying that at least for the case k even we would expect T_n to be small for $\zeta = -s, se^{i\theta_1}, se^{i\theta_2}, \dots, se^{i\theta(k-2)/2}$ where $s < 1$, but near 1. For the cases $k = 2, 3, \dots, 7$ if we choose

$$r_m = se^{2\pi im/k} \quad m = 1, 2, \dots, k-1$$

the β_i 's ≥ 0 for s sufficiently large because for $s = 1$ these roots generate the Newton-Cotes formulas where the β_i 's are positive for $k < 8$.

Degree	$\alpha - \max$	$\alpha - \min$
$k = 3$	$\beta \geq 0$	$\beta \geq 0$
$k = 4$	$\beta \geq 0$	$\beta \geq 0$
$k = 5$	$\beta \geq 0$	$\beta \geq 0$
$k = 6$	$\beta \geq 0$	$\beta \geq 0$
$k = 7$	$\beta \geq 0$	$\beta \geq 0$
$k = 8$	$\beta \geq 0$	$\beta \geq 0$
$k = 9$	$\beta \geq 0$	$\beta \geq 0$
$k = 10$	$\beta \geq 0$	$\beta < 0$
$k = 11$	$\beta \geq 0$	$\beta \geq 0$
$k = 12$	$\beta \geq 0$	$\beta < 0$
$k = 13$	$\beta \geq 0$	$\beta < 0$
$k = 14$	$\beta \geq 0$	$\beta < 0$
$k = 15$	$\beta \geq 0$	$\beta < 0$
$k = 16$	$\beta \geq 0$	$\beta < 0$
$k = 17$	$\beta \geq 0$	$\beta < 0$
$k = 18$	$\beta \geq 0$	$\beta < 0$

Table 1

CHAPTER V. EXTENSION TO HIGHER DIMENSIONS

Consider $\mu \frac{dy}{dx} = f(y, x)$ where

$$y(x) = \begin{vmatrix} y^1(x) \\ y^2(x) \\ \vdots \\ y^s(x) \end{vmatrix} \quad \text{and} \quad f(y, x) = \begin{vmatrix} f^1(y^1, y^2, \dots, y^s, x) \\ f^2(y^1, y^2, \dots, y^s, x) \\ \vdots \\ f^s(y^1, y^2, \dots, y^s, x) \end{vmatrix}.$$

We can use the same notation as previously provided we make the following definitions:

$$f_y(y, x) = \begin{vmatrix} \frac{\partial f^1(y, x)}{\partial y^1}, \frac{\partial f^1(y, x)}{\partial y^2}, \dots, \frac{\partial f^1(y, x)}{\partial y^s} \\ \frac{\partial f^2(y, x)}{\partial y^1}, \frac{\partial f^2(y, x)}{\partial y^2}, \dots, \frac{\partial f^2(y, x)}{\partial y^s} \\ \vdots \\ \frac{\partial f^s(y, x)}{\partial y^1}, \frac{\partial f^s(y, x)}{\partial y^2}, \dots, \frac{\partial f^s(y, x)}{\partial y^s} \end{vmatrix}.$$

Similar meanings will be prescribed to f_{xx} , f_{xy} , and f_{yy} . Also, we have

$$y_n = \begin{pmatrix} y^1(x_n) \\ y^2(x_n) \\ \vdots \\ y^s(x_n) \end{pmatrix}, \quad Y_n = \begin{pmatrix} Y_n^1 \\ Y_n^2 \\ \vdots \\ Y_n^s \end{pmatrix},$$

$$F_n = \begin{pmatrix} f^1(Y_n^1, Y_n^2, \dots, Y_n^s, x_n) \\ \vdots \\ f^s(Y_n^1, Y_n^2, \dots, Y_n^s, x_n) \end{pmatrix}, \quad f_n = f(y(x_n), x_n),$$

$$\phi(x_n) = \begin{pmatrix} \phi^1(x_n) \\ \phi^2(x_n) \\ \vdots \\ \phi^s(x_n) \end{pmatrix}, \quad \text{and} \quad f_{yn} = f_y|_{x=x_n}.$$

Finally, we replace the absolute value with the maximum norm:

$$|y_n| = \max_{i=1,2,\dots,s} |y_n^i| = \|y_n\|_\infty.$$

Let $A = (a_{ij})$ be an s by s matrix. Then

$$|A| = \|A\|_\infty = \max_i \sum_{k=1}^s |a_{ik}|.$$

Using this notation we consider the linear system defined by

$$\mu \frac{dy}{dx} = f(y,x) = \begin{pmatrix} a_{11}y^1 + a_{12}y^2 + \cdots + a_{1s}y^s \\ a_{21}y^1 + a_{22}y^2 + \cdots + a_{2s}y^s \\ \vdots \\ a_{s1}y^1 + a_{s2}y^2 + \cdots + a_{ss}y^s \end{pmatrix}$$

or $\mu \frac{dy}{dx} = Ay$ where $A = (a_{ij})$ is a constant matrix.

The related difference equation is

$$\begin{aligned} (5.1) \quad & \alpha_k Y_{n+k} + \alpha_{k-1} Y_{n+k-1} + \cdots + \alpha_0 Y_n \\ & = h^{1-\gamma} (\beta_k A Y_{n+k} + \beta_{k-1} A Y_{n+k-1} + \cdots + \beta_0 A Y_n). \end{aligned}$$

Let P be the matrix which reduces A to its Jordan Canonical form. Let $PV_n = Y_n$ and $P^{-1}AP = J$. It should be noted that P is a constant matrix. Substituting into equation (5.1) and multiplying on the left by P^{-1} , we get

$$\begin{aligned} & \alpha_k V_{n+k} + \alpha_{k-1} V_{n+k-1} + \cdots + \alpha_0 V_n \\ & = h^{1-\gamma} (\beta_k J V_{n+k} + \beta_{k-1} J V_{n+k-1} + \cdots + \beta_0 J V_n). \end{aligned}$$

First of all let us assume J is diagonal with eigenvalues λ_i ($i = 1, 2, \dots, s$).

We look at the component form of the last expression:

$$\alpha_k V_{n+k}^i + \dots + \alpha_o V_n^i = h^{1-\gamma} (\beta_k \lambda_i V_{n+k}^i + \dots + \beta_o \lambda_i V_n^i).$$

By Theorem 2.1 for h sufficiently small the roots of $\rho(\zeta) - h^{1-\gamma} \lambda_i \sigma(\zeta) = 0$ are of the form

$$(5.2) \quad \zeta_j = \zeta_{j0} + \frac{(m! \sigma(\zeta_{j0}) \lambda_i h^{1-\gamma})^{1/m_j}}{\rho^{(m_j)}(\zeta_{j0})} + \mathcal{O}(h^{2(1-\gamma)/m_j})$$

where $\rho(\zeta_{j0}) = 0$ and m_j is the multiplicity of ζ_{j0} .

However, we have imposed no conditions on λ_i as yet.

Since $\zeta_{j0} = 1$ is always one of the roots of $\rho(\zeta)$, we must first stipulate that $\operatorname{Re} \lambda_i < 0$. As in the one dimensional case if $|\zeta_{j0}| < 1$, then

$$\lim_{\substack{h \rightarrow 0 \\ x_n = nh}} |\zeta_j|^n = 0$$

Assume now $\zeta_{j0} = e^{i\theta}$ and hence $m = 1$. Let $\lambda_i = -\lambda_{i0} + i\lambda_{i1}$ where $\lambda_{i0} > 0$. Recall that

$$(5.3) \quad \frac{-\sigma(\zeta_{j0})}{\rho'(\zeta_{j0})} = p_j + iq_j.$$

Equation (5.2) may now be written as

$$(5.4) \quad \zeta_j = \cos\theta + (\lambda_{i0}p_j + \lambda_{i1}q_j)h^{1-\gamma} \\ + i[\sin\theta + (\lambda_{i0}q_j - \lambda_{i1}p_j)h^{1-\gamma}] + \mathcal{O}(h^{2(1-\gamma)})$$

and consequently

$$(5.5) \quad |\zeta_j|^2 = 1 + [2(\lambda_{i0}p_j + \lambda_{i1}q_j)\cos\theta \\ + 2(\lambda_{i0}q_j - \lambda_{i1}p_j)\sin\theta]h^{1-\gamma} + \dots$$

In order that $\lim_{\substack{h \rightarrow 0 \\ x_n = nh}} |\zeta_j|^n$ converge, we must require that

the term in brackets in (5.5) be negative. In general this is not a very practical condition, since it depends upon the eigenvalues of A , which are difficult quantities to obtain.

We may obtain a more practical condition by requiring that the eigenvalues be real and negative ($\lambda_{i1} = 0$) when we wish to consider difference schemes having roots of $\rho(\zeta)$ on the unit circle, but not identically equal to unity.

When $\lambda_{i1} = 0$ ($i = 1, 2, \dots, s$) we have

$$(5.6) \quad |\zeta_j|^2 = 1 + [2(\lambda_{i0}p_j\cos\theta + \lambda_{i0}q_j\sin\theta)]h^{1-\gamma} + \dots$$

Our condition now becomes

$$(5.7) \quad p_j \cos \theta + q_j \sin \theta < 0;$$

that is, we require μ -stability in order to guarantee convergence.

The following theorem will be useful in order to assure that the eigenvalues of A are real and negative:

THEOREM: Suppose that A is symmetric, $a_{ii} < 0$ and

$$a_{ii} + \sum_{j=1, j \neq i}^s |a_{ij}| \leq -M < 0 \quad i = 1, 2, \dots, s$$

where the prime on the summation sign denotes $|a_{ii}|$ is not included in the sum. Then the eigenvalues of A are real and strictly negative.

Proof: See, for example, Varga: Matrix Iterative Analysis.

We give an example of a scheme which is μ -stable but which does not give convergent results because the term in brackets in (5.5) for this example is positive, i.e., this method can be used for the scalar boundary layer equation, but not the vector one.

Example 1. Let $\rho(\zeta) = \zeta^3 - \zeta^2 + \zeta - 1$. The roots are $\zeta = 1, \pm i$, and the μ -stability condition implies that

$$1 > \beta_1 + \beta_2.$$

Let $\sigma(\zeta) = \frac{1}{2} \zeta^3 + \frac{1}{2} \zeta + 1$. Therefore,

$$\frac{-\sigma(i)}{\rho'(i)} = \frac{1}{4} - \frac{1}{4} i.$$

Suppose the Jordan Canonical form of A is

$$J = \begin{pmatrix} -2-4i & 0 \\ 0 & -2+4i \end{pmatrix}.$$

Substituting the eigenvalue $-2-4i$ and $\zeta_{j0} = i$ into (5.5), we obtain

$$|\zeta_j|^2 = 1 + h^{1-\gamma} + O(h^{2(1-\gamma)}).$$

In order to prove the analogy of Theorem 4.1 for the vector equation $\mu y' = f(y, x)$ we have seen that in certain cases it is necessary to restrict the class of acceptable functions $f(y, x)$. First, as before, we assume $f_y(y(x), x)$ is continuous and bounded in D . We distinguish between two cases:

(5.8) Case 1. The roots of $\rho(\zeta)$ lie inside the unit circle with the exception of $\zeta = 1$. Here we require that the real parts of the roots of the characteristic equation

$$\text{Det} \parallel \frac{\partial f(y(x, \mu), x)}{\partial y} - \lambda I \parallel = 0$$

are negative and bounded away from zero in D for all μ sufficiently small.

(5.9) Case 2. The roots of $\rho(\xi)$ other than $\xi = 1$ may lie on the unit circle. Here in addition to μ -stability we must require that the roots of the characteristic equation are strictly negative and bounded away from zero for all values of x in D and μ sufficiently small.

Let us now investigate the form of Duhamel's principle for the consistent difference equation satisfying condition (5.8) or (5.9)

$$(5.10) \quad \sum_{i=0}^k \alpha_i Y_{n+i} - h^{1-\gamma} \sum_{i=0}^k \beta_i J Y_{n+i} = Q_n.$$

J is a matrix in Jordan Canonical form. We claim that the solution to (5.10) takes the form

$$(5.11) \quad Y_n = \begin{cases} \sum_{j=0}^{n-k} \bar{\Phi}_{n-j-1} Q_j + \Psi_n & n \geq k \\ \Psi_n & n < k \end{cases}$$

where $\bar{\Phi}_n$ is defined as a matrix satisfying the following equation:

$$(5.12) \quad \sum_{i=0}^k \bar{\Phi}_{n+i} (\alpha_i - h^{1-\gamma} \beta_i J) = 0$$

with initial conditions $\bar{\Phi}_0 = \bar{\Phi}_1 = \dots = \bar{\Phi}_{k-2} \equiv 0$ and $\bar{\Phi}_{k-1}$ is chosen to satisfy the equation

$$(5.13) \quad \bar{\Phi}_{k-1} (\alpha_k - h^{1-\gamma} \beta_k J) = \begin{vmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{vmatrix} \equiv 1$$

Since J is in canonical form we see immediately that

$$(5.14) \quad \begin{aligned} \bar{\Phi}_{k-1}^{ij} &= 0 \quad \text{if } i > j \\ \bar{\Phi}_{k-1}^{ii} &= (\alpha_k - h^{1-\gamma} \beta_k J)^{-1}. \end{aligned}$$

The other elements of the matrix $\bar{\Phi}_{k-1}$ will depend upon the particular form of J .

Ψ_n is the solution to the homogeneous equation ((5.10) with $Q_n = 0$) with initial data $Y_0, Y_1, \dots, Y_{k-1}$. Therefore,

$$(5.15) \quad \Psi_n = \sum_{\ell=0}^{k-1} \left(\sum_{j=0}^{k-\ell-1} (\alpha_{k-j} - h^{1-\gamma} \beta_{k-j} J) \bar{\Phi}_{n+k-\ell-j-1} \right) Y_\ell$$

for $n \geq k$

where $\bar{\Phi}_n$ is defined by

$$(5.16) \quad \sum_{i=0}^k (\alpha_i - h^{1-\gamma} \beta_i^J) \bar{\Phi}_{n+i} = 0$$

with initial conditions $\bar{\Phi}_0 = \bar{\Phi}_1 = \dots = \bar{\Phi}_{k-2} = 0$ and $\bar{\Phi}_{k-1}$ is chosen to satisfy the equation

$$(5.17) \quad (\alpha_k - h^{1-\gamma} \beta_k^J) \bar{\Phi}_{k-1} = 1.$$

To verify that (5.11) is really the solution to (5.10) we consider the homogeneous solution, Ψ_n , first:

$$\begin{aligned} & \sum_{i=0}^k (\alpha_i - h^{1-\gamma} \beta_i^J) \Psi_{n+i} \\ &= \sum_{i=0}^k (\alpha_i - h^{1-\gamma} \beta_i^J) \sum_{\ell=0}^{k-1} \left(\sum_{j=0}^{k-\ell-1} (\alpha_{k-j} - h^{1-\gamma} \beta_{k-j}^J) \right. \\ & \quad \left. \bar{\Phi}_{n+i+k-\ell-j-1} \right) Y_\ell \\ &= \sum_{\ell=0}^{k-1} \left(\sum_{j=0}^{k-\ell-1} \sum_{i=0}^k (\alpha_{k-j} - h^{1-\gamma} \beta_{k-j}^J) (\alpha_i - h^{1-\gamma} \beta_i^J) \right. \\ & \quad \left. \bar{\Phi}_{n+i+k-\ell-j-1} \right) Y_\ell \\ &= 0 \quad \text{for} \quad n \geq 0 \end{aligned}$$

$$\text{because} \quad \sum_{i=0}^k (\alpha_i - h^{1-\gamma} \beta_i^J) \bar{\Phi}_{n+k+i-\ell-j-1} = 0 \quad \text{for} \quad n \geq 0$$

and $0 \leq \ell + j \leq k - 1$.

For $n < k$, we have

$$\begin{aligned}
 \Psi_n &= \sum_{\ell=0}^{k-1} \left(\sum_{j=0}^{k-\ell-1} (\alpha_{k-j} - h^{1-\gamma} \beta_{k-j}^J) \bar{\Phi}_{n+k-\ell-j-1} \right) Y_\ell \\
 &= (\alpha_k - h^{1-\gamma} \beta_k^J) \bar{\Phi}_{k-1} Y_n \\
 &= Y_n .
 \end{aligned}$$

On the other hand, the particular solution satisfies the equation (5.10) because

$$\begin{aligned}
 \sum_{j=0}^{n-k} \bar{\Phi}_{n-j-1} Q_j &= \sum_{j=0}^{n-k} \bar{\Phi}_{n-j-1} \left(\sum_{i=0}^k (\alpha_i - h^{1-\gamma} \beta_i^J) Y_{j+i} \right) \\
 &= \sum_{i=0}^k \sum_{j=0}^{n-k} \bar{\Phi}_{n-j-1} (\alpha_i - h^{1-\gamma} \beta_i^J) Y_{j+i} .
 \end{aligned}$$

Let $j + i = \ell$. Then

$$\sum_{j=0}^{n-k} \bar{\Phi}_{n-j-1} Q_j = \sum_{i=0}^k \sum_{\ell=i}^{n+i-k} \bar{\Phi}_{n-\ell-1+i} (\alpha_i - h^{1-\gamma} \beta_i^J) Y_\ell .$$

But for the particular solution we impose the initial values $Y_\ell = 0$ for $\ell < k$. Since initially $\bar{\Phi}_0 = \bar{\Phi}_1 = \dots = \bar{\Phi}_{k-2} = 0$, we have

$$\sum_{j=0}^{n-k} \bar{\Phi}_{n-j-1} Q_j = \sum_{i=0}^k \sum_{\ell=0}^{n+i} \bar{\Phi}_{n-\ell+i-1} (\alpha_i - h^{1-\gamma} \beta_i^J) Y_\ell$$

$$= \sum_{i=0}^k \sum_{\ell=0}^n \bar{\Phi}_{n-\ell+i-1} (\alpha_i - h^{1-\gamma} \beta_i J) Y_\ell$$

where the last result follows from the definition of

$$\bar{\Phi}_{-n} \equiv 0 \quad \text{for } n = 1, 2, \dots$$

Hence,

$$\begin{aligned} \sum_{j=0}^{n-k} \bar{\Phi}_{n-j-1} Q_j &= \sum_{\ell=0}^n \sum_{i=0}^k \bar{\Phi}_{n-\ell+i-1} (\alpha_i - h^{1-\gamma} \beta_i J) Y_\ell \\ &= \sum_{i=0}^k \bar{\Phi}_{-1+i} (\alpha_i - h^{1-\gamma} \beta_i J) Y_n \\ &= \bar{\Phi}_{k-1} (\alpha_k - h^{1-\gamma} \beta_k J) Y_n \\ &= Y_n \quad \text{by (5.12) and (5.13).} \end{aligned}$$

It is now necessary to prove lemmas analogous to Lemmas 1 and 2. We will merely illustrate the proof for a special case; namely,

$$(5.18) \quad J = \begin{vmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 \\ 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & \lambda_2 \end{vmatrix}.$$

The more general Jordan form lends itself to a similar proof.

LEMMA 1'. Let the consistent difference equation

$$(5.19) \quad \sum_{i=0}^k \bar{\Phi}_{n+i} (\alpha_i - h^{1-\gamma} \beta_i J) = 0$$

satisfy condition (5.8) or (5.9) and initial conditions

$$\bar{\Phi}_0 = \bar{\Phi}_1 = \dots = \bar{\Phi}_{k-2} = 0 \quad \text{and} \quad \bar{\Phi}_{k-1} = (\alpha_k - h^{1-\gamma} \beta_k J)^{-1}.$$

Then

$$\sum_{i=0}^N |\bar{\Phi}_i^{jk}| \leq \frac{C}{h^{1-\gamma}} \quad j, k = 1, 2, \dots, s$$

for h sufficiently small and N any integer.

Proof: We supply the proof for the case when J has the form given by (5.18). The initial conditions on $\bar{\Phi}_{k-1}$ are given by

$$\bar{\Phi}_{k-1} (\alpha_k - h^{1-\gamma} \beta_k J) = 1.$$

Equating terms gives

$$\bar{\Phi}_{k-1}^{11} = (\alpha_k - h^{1-\gamma} \beta_k \lambda_1)^{-1} \quad \bar{\Phi}_{k-1}^{21} = \bar{\Phi}_{k-1}^{31} = \bar{\Phi}_{k-1}^{41} = 0$$

$$\bar{\Phi}_{k-1}^{12} = \frac{h^{1-\gamma} \beta_k}{(\alpha_k - h^{1-\gamma} \lambda_1 \beta_k)^2} \quad \bar{\Phi}_{k-1}^{22} = (\alpha_k - h^{1-\gamma} \beta_k \lambda_1)^{-1} \quad \bar{\Phi}_{k-1}^{32} = \bar{\Phi}_{k-1}^{42} = 0$$

$$\bar{\Phi}_{k-1}^{13} = \frac{(h^{1-\gamma} \beta_k)^2}{(\alpha_k - h^{1-\gamma} \lambda_1 \beta_k)^3} \quad \bar{\Phi}_{k-1}^{23} = \frac{h^{1-\gamma} \beta_k}{(\alpha_k - h^{1-\gamma} \lambda_1 \beta_k)^2}$$

$$\overline{\Phi}_{k-1}^{33} = (\alpha_k - h^{1-\gamma} \beta_k \lambda_1)^{-1} \quad \overline{\Phi}_{k-1}^{43} = 0$$

$$\overline{\Phi}_{k-1}^{14} = \overline{\Phi}_{k-1}^{24} = \overline{\Phi}_{k-1}^{34} = 0 \quad \overline{\Phi}_{k-1}^{44} = (\alpha_k - h^{1-\gamma} \beta_k \lambda_2)^{-1}.$$

With these initial conditions it is clear that

$$\overline{\Phi}_n^{ij} = 0 \quad \text{for } i > j \text{ and } n = 0, 1, 2, \dots$$

$$\text{Also, } \overline{\Phi}_n^{14} = \overline{\Phi}_n^{24} = \overline{\Phi}_n^{34} = 0 \quad n = 0, 1, 2, \dots$$

By Lemma 1 and conditions (5.8) or (5.9) on λ_1 and λ_2 it follows that

$$\sum_{i=0}^{n-1} |\overline{\Phi}_i^{jj}| \leq \frac{C}{h^{1-\gamma}} \quad j = 1, 2, \dots, s.$$

The remaining elements of $\overline{\Phi}$ satisfy the following equations:

$$a) \quad \sum_{i=0}^k (\alpha_i - h^{1-\gamma} \beta_i \lambda_1) \overline{\Phi}_{n+i}^{12} = h^{1-\gamma} \sum_{i=0}^k \beta_i \overline{\Phi}_{n+i}^{11}$$

$$b) \quad \sum_{i=0}^k (\alpha_i - h^{1-\gamma} \beta_i \lambda_1) \overline{\Phi}_{n+i}^{13} = h^{1-\gamma} \sum_{i=0}^k \beta_i \overline{\Phi}_{n+i}^{12}$$

$$c) \quad \sum_{i=0}^k (\alpha_i - h^{1-\gamma} \beta_i \lambda_1) \overline{\Phi}_{n+i}^{23} = h^{1-\gamma} \sum_{i=0}^k \beta_i \overline{\Phi}_{n+i}^{22}.$$

We solve a) by Duhamel's principle:

$$\begin{aligned}
 \bar{\Phi}_n^{12} &= \sum_{j=0}^{n-k} \hat{\Phi}_{n-j-1} (h^{1-\gamma} \sum_{i=0}^k \beta_i \bar{\Phi}_{j+i}^{11}) + \hat{\Psi}_n & n \geq k \\
 &= \hat{\Psi}_n & n < k
 \end{aligned}
 \tag{5.20}$$

where $\hat{\Phi}_n$ the one dimensional Green's function defined in Lemma 1. We need the following theorem from advanced calculus:

THEOREM. Assume that $\sum_{n=0}^{\infty} a_n$ converges absolutely and has sum A, and suppose $\sum_{n=0}^{\infty} b_n$ converges with sum B. Then $\sum_{n=0}^{\infty} c_n$ where $c_n = \sum_{k=0}^{\infty} a_k b_{n-k}$ converges and has sum AB.

Proof: See Mathematical Analysis by Apostol.

From (5.20) we have

$$\begin{aligned}
 |\bar{\Phi}_n^{12}| &\leq h^{1-\gamma} \sum_{j=0}^{n-k} |\hat{\Phi}_{n-j-1}| [|\beta_0| |\bar{\Phi}_j^{11}| + \dots + |\beta_k| |\bar{\Phi}_{j+k}^{11}|] + |\hat{\Psi}_n| \\
 &\leq h^{1-\gamma} \sum_{j=0}^{n-1} |\hat{\Phi}_{n-j-1}| [|\beta_0| + |\beta_1| + \dots + |\beta_k|] |\bar{\Phi}_j^{11}| + |\hat{\Psi}_n|
 \end{aligned}
 \tag{5.21}$$

since $|\bar{\Phi}_j^{11}|$ is decaying.

$$\sum_{n=0}^N |\bar{\Phi}_n^{12}| \leq h^{1-\gamma} \frac{C}{h^{1-\gamma}} \frac{C}{h^{1-\gamma}} + \sum_{n=0}^N |\hat{\Psi}_n| \leq \frac{C}{h^{1-\gamma}} + \sum_{n=0}^N |\hat{\Psi}_n|$$

where we have made use of Lemma 1 on the functions $\hat{\Phi}_n$ and $\hat{\Phi}_n^{11}$ and the above stated theorem. Now

$$\hat{\Psi}_n = \hat{\Phi}_n^{12} \quad n < k$$

and

$$\begin{aligned} \hat{\Psi}_n &= \sum_{i=0}^{k-1} \left(\sum_{j=0}^{k-i-1} (\alpha_{k-j} - h^{1-\gamma} \beta_{k-j} \lambda_1) \hat{\Phi}_{n+k-i-j-1} \right) \bar{\Phi}_i^{12} \\ &= (\alpha_k - h^{1-\gamma} \beta_k \lambda_1) \hat{\Phi}_n \bar{\Phi}_{k-1}^{12} \\ &= \frac{h^{1-\gamma} \beta_k}{(\alpha_k - h^{1-\gamma} \beta_k \lambda_1)} \hat{\Phi}_n \quad n \geq k \end{aligned}$$

where we have used the initial conditions for $\bar{\Phi}_i^{12}$.

Therefore,

$$\begin{aligned} \sum_{n=0}^N |\hat{\Psi}_n| &\leq |\bar{\Phi}_{k-1}^{12}| + \frac{\beta h^{1-\gamma}}{|\alpha_k - h^{1-\gamma} \beta_k \lambda_1|} \sum_{n=k}^N |\hat{\Phi}_n| \\ &\leq C \quad \text{by Lemma 1 again.} \end{aligned}$$

We therefore have that

$$\sum_{n=0}^N |\bar{\Phi}_n^{12}| \leq \frac{C}{h^{1-\gamma}}.$$

Equations b) and c) are treated in the same way. Lemma 1' is now complete.

LEMMA 2'.

$$|\bar{\Phi}_n^{jk}| \leq \left[\frac{C}{h^{(1-\gamma)(m-1)/m}} r^n n^{s-1} + C(1 - h^{1-\gamma_L})^n n^{s-1} \right]$$

for all n and $j, k = 1, 2, \dots, s$.

Proof: For $|\bar{\Phi}_n^{ii}|$ the proof is the same as Lemma 2. From (5.21) and Lemma 2

$$\begin{aligned} |\bar{\Phi}_n^{12}| &\leq h^{1-\gamma_\beta} \sum_{j=0}^{n-k} |\hat{\Phi}_{n-j-1}| (|\bar{\Phi}_j^{11}| + \dots + |\bar{\Phi}_{j+k}^{11}|) + |\hat{\Psi}_n| \\ &\leq (k+1)h^{1-\gamma_\beta} \sum_{j=0}^{n-k} \left[\frac{Cr^{n-j}}{h^{1-\gamma(m-1)/m}} + C(1-h^{1-\gamma_L})^{n-j} \right] \\ &\quad \left(\frac{Cr^j}{h^{(1-\gamma)(m-1)/m}} + C(1-h^{1-\gamma_L})^j \right) + |\hat{\Psi}_n| \\ &\leq \frac{n Cr^n}{h^{(1-\gamma)(m-1)/m}} + nC(1 - h^{1-\gamma_L})^n \end{aligned}$$

where we recall that $\hat{\Psi}_n = \frac{h^{1-\gamma_\beta k}}{\alpha_k - h^{1-\gamma_\beta k \lambda_1}} \hat{\Phi}_n$.

For $\bar{\Phi}_n^{13}$ we have

$$\begin{aligned}
|\bar{\Phi}_n^{13}| &\leq h^{1-\gamma_\beta} \sum_{j=0}^{n-k} |\hat{\Phi}_{n-j-1}| (|\bar{\Phi}_j^{12}| + \dots + |\bar{\Phi}_{j+k}^{12}|) + |\hat{\Psi}_n| \\
&\leq \frac{n^2 C r^n}{h^{1-\gamma(m-1)/m}} + n^2 C (1 - h^{1-\gamma_L})^n.
\end{aligned}$$

LEMMA 2".

$$|\bar{\Phi}_n^{jk}| \leq \frac{n^{s-1} r^n C}{h^{(1-\gamma)(m-1)/m}} + n^{s-1} (1 - h^{1-\gamma_L})^n C$$

for all $n, j, k = 1, 2, \dots, s$, $0 < r < 1$, $L > 0$ and $m = \max\{m_1, m_2, \dots, m_\ell\}$.

Proof: The proof follows the same kind of argument as the ones found in Lemma 1' and Lemma 2'.

With the aid of these lemmas we can prove the following theorem:

THEOREM 5.1. Let the consistent finite difference equation

$$\sum_{i=0}^k \alpha_i Y_{n+i} = h^{1-\gamma} \sum_{i=0}^k \beta_i A Y_{n+i} + R_n$$

where $A = (a_{ij})$ is a constant matrix, satisfy the following conditions:

- a) $R_n = O(h^{2-2\gamma})$ as $h \rightarrow 0$
 b) Condition (5.8) or condition (5.9) is satisfied
 c) $|Y_i - y(x_i)| \leq Th^{1-\gamma}$ for h sufficiently small
 and $i = 0, 1, \dots, k-1$. T is a positive constant independent of h .

Then the difference equation is convergent.

Proof: Subtracting the difference equation satisfied by y_n from the above and setting $e_n = Y_n - y_n$ gives

$$\sum_{i=0}^k \alpha_i e_{n+i} = h^{1-\gamma} \sum_{i=0}^k \beta_i A e_{n+i} + R_n - T_n.$$

If we let $e_n = Pu_n$ where $P^{-1}AP = J$ and multiply on the left by P^{-1} , we obtain

$$\sum_{i=0}^k \alpha_i u_{n+i} = h^{1-\gamma} \sum_{i=0}^k \beta_i J u_{n+i} + P^{-1}(R_n - T_n).$$

The proof now follows the same course as Theorem 2.2 with the use of Lemmas 1', 2' and 2" instead of Lemmas 1 and 2. However, we must choose $N_0 = s \frac{\ln h}{\ln r}$ instead of $\frac{\ln h^{1-\gamma}}{\ln r}$ because of the slight difference between Lemma 2" and Lemma 2. This becomes clear when we note that

$$\frac{n^{s-1} r^n}{h^{(1-\gamma)(m-1)/m}} \leq \frac{x_n^{s-1} r^n}{h^s} \leq x_n^{s-1} \leq 1$$

and

$$\lim_{h \rightarrow 0} h^{1-\gamma} \ln h = \lim_{h \rightarrow 0} \frac{1}{\gamma-1} h^{1-\gamma} = 0.$$

Q.E.D.

THEOREM 5.2. Let the consistent finite difference equation

$$(5.22) \quad \sum_{i=0}^k \alpha_i Y_{n+i} = h^{1-\gamma} \sum_{i=0}^k \beta_i F_{n+i} + R_n$$

where

$$F_n = \begin{pmatrix} f^1(Y_n^1, Y_n^2, \dots, Y_n^s, x_n) \\ f^2(Y_n^1, Y_n^2, \dots, Y_n^s, x_n) \\ \vdots \\ f^s(Y_n^1, Y_n^2, \dots, Y_n^s, x_n) \end{pmatrix}$$

satisfy the following conditions:

a) $R_n = O(h^{2(1-\gamma)})$ as $h \rightarrow 0$

b) The matrices $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, and $\frac{\partial^2 f}{\partial y^2}$ are continuous and

bounded for $0 \leq x \leq 1$ and $-\infty < y^i < \infty$ $i = 1, 2, \dots, s$

c) Condition (5.8) or condition (5.9) is satisfied.

d) $|e_i| \leq Th^{1-\gamma}$ for h sufficiently small and $i = 0, 1, \dots, k-1$. T is a positive constant.

Then there exists a constant C such that $|e_n| \leq Ch^{1-\gamma}$ for $n = 0, 1, \dots, N$ where $x_N = 1$.

Proof: The exact solution of $\mu y' = f(y, x)$ satisfies the difference equation

$$(5.23) \quad \sum_{i=0}^k \alpha_i y_{n+i} = h^{1-\gamma} \sum_{i=0}^k \beta_i f_{n+i} + T_n$$

where T_n can be shown to be $\mathcal{O}(h^{2-2\gamma})$ as $h \rightarrow 0$ in the same way as in Theorem 4.1.

Subtracting (5.23) and (5.22) and letting $e_i = Y_i - y_i$, we obtain

$$\sum_{i=0}^k \alpha_i e_{n+i} = h^{1-\gamma} \sum_{i=0}^k \beta_i (F_{n+i} - f_{n+i}) + R_n - T_n.$$

In component form we can write

$$\sum_{i=0}^k \alpha_i e_{n+i}^{\ell} = h^{1-\gamma} \sum_{i=0}^k \beta_i (F_{n+i}^{\ell} - f_{n+i}^{\ell}) + R_n^{\ell} - T_n^{\ell}$$

for $\ell = 1, 2, \dots, s$.

$$= h^{1-\gamma} \sum_{i=0}^k \beta_i \sum_{j=1}^s \frac{\partial \bar{f}_{n+i}^{\ell}}{\partial y^j} e_{n+i}^j + R_n^{\ell} - T_n^{\ell}$$

where

$$\frac{\partial \bar{f}_{n+i}^{\ell}}{\partial y^j} = \frac{\partial f^{\ell}}{\partial y^j}(y_{n+i}^1 + \theta_{n+i}^{1\ell} e_{n+i}^1, \dots, y_{n+i}^s + \theta_{n+i}^{s\ell} e_{n+i}^s, x_{n+i})$$

$\ell = 1, 2, \dots, s$

$j = 1, 2, \dots, s$

Therefore, in vector notation we have

$$(5.24) \quad \sum_{i=0}^k \alpha_i e_{n+i} = h^{1-\gamma} \sum_{i=0}^k \beta_i \frac{\partial \bar{f}}{\partial y} e_{n+i} + R_n - T_n$$

With the choice of $N_0 = s \frac{\ell n}{\ell n} \frac{h}{r}$ we can find a bound for $|e_n|$ for $0 \leq n \leq N_0$ as was done in Theorem 4.1. Assuming this has been accomplished, we add

$$-h^{1-\gamma} \sum_{i=0}^k \beta_i L_0 e_{n+i} = -h^{1-\gamma} \sum_{i=0}^k \beta_i \frac{\partial f(y(x_0), x_0)}{\partial y} e_{n+i}$$

to both sides of (5.24) and make the substitution $e_n = Pu_n$ where $P^{-1}L_0P = J$, the Jordan Canonical form of L_0 . After multiplication by P^{-1} , we have

$$(5.25) \quad \sum_{i=0}^k (\alpha_i - h^{1-\gamma} \beta_i J) u_{n+i} =$$

$$h^{1-\gamma} \sum_{i=0}^k \beta_i P^{-1} \left(\frac{\partial \bar{f}}{\partial y} e_{n+i} - \frac{\partial f}{\partial y} o \right) Pu_{n+i} + P^{-1}(R_n - T_n).$$

We let Q_n denote the right side of (5.25).

Using condition b) of the hypothesis we can write

$$\left| \frac{\partial \bar{f}^\ell}{\partial y^j} e_{n+i} - \frac{\partial f^\ell}{\partial y^j} o \right| \leq \left| \frac{\partial f^\ell}{\partial y^j} e_{n+i} - \frac{\partial f^\ell}{\partial y^j} o \right| + O(|e_{n+i}|)$$

$$\leq \varepsilon + O(|u_{n+i}|)$$

where the ε can be chosen as small as necessary by continuity.

A bound for Q_n is then

$$|Q_n| \leq h^{1-\gamma} \sum_{i=0}^k |P^{-1}| |P| (|\beta_i| s \varepsilon) |u_{n+i}| \\ + \mathcal{O}(h^{2(1-\gamma)}) + \mathcal{O}(h^{1-\gamma} \sum_{i=0}^k |u_{n+i}|^2)$$

where $\mathcal{O}(h^{2(1-\gamma)})$ is due to the round-off and truncation errors.

The solution to the system (5.25) was given earlier as

$$(5.26) \quad u_n = \begin{cases} \sum_{j=0}^{n-k} \bar{\Phi}_{n-j-1} Q_j + \Psi_n & n \geq k \\ \Psi_n & n < k \end{cases}$$

If we now introduce the max norm $U_n = \max_{i=0,1,\dots,n} |u_i|$ and

follow the same procedure as in Theorem 4.1 but using Lemmas 1', 2', and 2'' instead of Lemmas 1 and 2, we obtain the same basic inequality

$$U_n \leq \frac{K_6 U_n^2}{1-a} + \frac{K_5 h^{1-\gamma}}{1-a} + \frac{K_4 T h^{1-\gamma}}{1-a} \quad 0 \leq n \leq N_1$$

where N_1 is determined by the choice of ε and continuity considerations.

The proof is completed exactly like Theorem 4.1. Q.E.D.

Since the estimates used in Theorem 5.2 are rather complex, it is not a simple matter to estimate the number of steps ℓ occurring in Theorem 5.2. Instead we show how better error bounds can be obtained when we consider a very restricted class of difference schemes.

THEOREM 5.3. Let the consistent explicit difference equation

$$\sum_{i=0}^k \alpha_i Y_{n+i} = h^{1-\gamma} \sum_{i=0}^{k-1} \beta_i F_{n+i} + R_n$$

satisfy the following conditions:

- a) $|R_n - T_n| \leq K_1 h^{(p+1)(1-\gamma)}$
- b) $e_i = O(h^{p(1-\gamma)})$ for $i = 0, 1, \dots, k-1$
- c) The matrix $\frac{\partial f}{\partial y}$ is continuous, bounded and strictly negatively diagonally dominate for all points of interest, i.e., if $A = (a_{ij})$, then

$$a_{ii} + \sum_{\substack{j=1 \\ i \neq j}}^s |a_{ij}| < 0 \quad \text{for all } i.$$

- d) $\alpha_k = 1, -\alpha_{k-1}, -\alpha_{k-2}, \dots, -\alpha_1, -\alpha_0 > 0$ and $\beta_{k-1}, \beta_{k-2}, \dots, \beta_0 \geq 0$.

Then the difference equation is convergent.

Proof: If we subtract the difference equation satisfied by y_n from the above and set $e_n = Y_n - y_n$, we obtain

$$\begin{aligned} e_{n+k} &= \sum_{i=0}^{k-1} [-\alpha_i e_{n+i} + h^{1-\gamma} \beta_i (F_{n+i} - f_{n+i})] + R_n - T_n \\ &= \sum_{i=0}^{k-1} [-\alpha_i + h^{1-\gamma} \beta_i \frac{\partial \bar{f}}{\partial y}] e_{n+i} + R_n - T_n \end{aligned}$$

where

$$\frac{\partial \bar{f}}{\partial y} = \begin{vmatrix} \frac{\partial f^1}{\partial y^1}(\zeta_{n+i}^1, x_{n+i}), \dots, \frac{\partial f^1}{\partial y^s}(\zeta_{n+i}^1, x_{n+i}) \\ \vdots \\ \frac{\partial f^s}{\partial y^1}(\zeta_{n+i}^s, x_{n+i}), \dots, \frac{\partial f^s}{\partial y^s}(\zeta_{n+i}^s, x_{n+i}) \end{vmatrix}$$

By condition c) there exists a number $M > 0$ such that for all numbers $a > 0$ and $b \geq 0$ and h sufficiently small

$$|a| + b h^{1-\gamma} \left| \frac{\partial \bar{f}}{\partial y} \right| \leq a - b M h^{1-\gamma}$$

for $n = 0, 1, \dots, N$
and $i = 0, 1, \dots, k$.

Taking norms and using the last inequality we obtain

$$|e_{n+k}| \leq \sum_{i=0}^{k-1} (-\alpha_i - \beta_i M h^{1-\gamma}) |e_{n+i}| + K_1 h^{(p+1)(1-\gamma)}$$

where conditions a) and d) were used.

We now prove by induction that

$$|e_{n+k}| \leq \frac{K_1 h^{p(1-\gamma)}}{M \sum_{i=0}^{k-1} \beta_i}$$

We assume this last inequality is satisfied for $|e_0|$, $|e_1|, \dots, |e_{k-1}|$. If it is not, choose a larger K_1 . By induction then we assume the inequality holds for $|e_k|$, $|e_{k+1}|, \dots, |e_{n+k-1}|$, then

$$\begin{aligned} |e_{n+k}| &\leq \sum_{i=0}^{k-1} (-\alpha_i - \beta_i M h^{1-\gamma}) \frac{K_1 h^{p(1-\gamma)}}{M \sum_{i=0}^{k-1} \beta_i} + K_1 h^{(p+1)(1-\gamma)} \\ &= \frac{K_1 h^{p(1-\gamma)}}{M \sum_{i=0}^{k-1} \beta_i} - K_1 h^{(p+1)(1-\gamma)} + K_1 h^{(p+1)(1-\gamma)} \\ &= \frac{K_1 h^{p(1-\gamma)}}{M \sum_{i=0}^{k-1} \beta_i} \end{aligned}$$

where we have used the consistency condition $\sum_{i=0}^{k-1} (-\alpha_i) = 1$.

Thus the proof is complete by induction.

VI. BOUNDARY LAYER EQUATIONS OF HIGHER ORDER

Suppose we wish to consider the second order equation

$$(6.1) \quad \mu \frac{d^2 z}{dx^2} + a(x) \frac{dz}{dx} + b(x) z = r(x).$$

$$\text{Let } \frac{dz}{dx} = y.$$

Equation (6.1) is equivalent to the system

$$\mu \frac{dy}{dx} = -a(x) y - b(x) z + r(x)$$

$$\frac{dz}{dx} = y.$$

We are therefore led in a natural way to investigate the system

$$(6.2) \quad \begin{aligned} \mu \frac{dy}{dx} &= f(y, z, x) \\ \frac{dz}{dx} &= g(y, z, x) \end{aligned} \quad 0 \leq x \leq 1$$

with initial conditions

$$y(0) = y^0$$

$$z(0) = z^0$$

Following our previous approach, we first consider the linear case

$$(6.3) \quad \mu \frac{dy}{dx} = a_{11}y + a_{12}z$$

$$\frac{dz}{dx} = a_{21}y + a_{22}z$$

where (a_{ij}) is a constant matrix. From the ODE theory we must require that $\frac{\partial f}{\partial y} = a_{11} < 0$.

Dividing the first of the above two equations by μ and using vector notation, we obtain

$$\begin{pmatrix} \frac{dy}{dx} \\ \frac{dz}{dx} \end{pmatrix} = A_\mu \begin{pmatrix} y \\ z \end{pmatrix}$$

where

$$A_\mu \equiv \begin{pmatrix} \frac{a_{11}}{\mu} & \frac{a_{12}}{\mu} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} \frac{a_{11}}{h^\gamma} & \frac{a_{12}}{h^\gamma} \\ a_{21} & a_{22} \end{pmatrix} \equiv A_h .$$

The characteristic equation of A_μ is

$$\left(\frac{a_{11}}{\mu} - \lambda\right)(a_{22} - \lambda) - a_{21} \frac{a_{12}}{\mu} = 0$$

and

$$(6.5) \quad \lambda = \frac{\mu a_{22} + a_{11} \pm \sqrt{(\mu a_{22} + a_{11})^2 - 4\mu \det A}}{2\mu}$$

where

$$A = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

$$(6.6) \quad \lambda_1 \sim \frac{a_{11}}{\mu} \text{ as } \mu \rightarrow 0$$

$$(6.7) \quad \lambda_2 = \frac{\mu a_{22} + a_{11} - [\mu a_{22} + a_{11} - 2\mu \det A (\mu a_{22} + a_{11})^{-1} - \frac{1}{8}(4\mu \det A)^2 (\mu a_{22} + a_{11})^{-3} + \dots]}{2\mu}$$

$$(6.8) \quad \lambda_2 \sim \frac{\det A}{a_{11}} \text{ as } \mu \rightarrow 0.$$

We note that for μ sufficiently small the eigenvalues are distinct and λ_2 is bounded.

The eigenvectors are found as follows:

$$A_{\mu} \begin{Bmatrix} \xi_1 \\ \xi_2 \end{Bmatrix} = \lambda_1 \begin{Bmatrix} \xi_1 \\ \xi_2 \end{Bmatrix}$$

$$\frac{a_{11}}{\mu} \xi_1 + \frac{a_{12}}{\mu} \xi_2 = \lambda_1 \xi_1$$

$$a_{21} \xi_1 + a_{22} \xi_2 = \lambda_1 \xi_2$$

$$a_{21} \xi_1 + a_{22} \xi_2 = \xi_2 \left[\frac{\mu a_{22} + a_{11} + \sqrt{(\mu a_{22} + a_{11})^2 - 4\mu \det A}}{2\mu} \right]$$

$$\frac{\xi_2}{\xi_1} = \frac{2\mu a_{21}}{[a_{11} - \mu a_{22} + \sqrt{(\mu a_{22} + a_{11})^2 - 4\mu \det A}]}$$

Letting $\xi_1 = 1$ we obtain the eigenvector $\begin{Bmatrix} 1 \\ \xi_2 \end{Bmatrix}$.

Remark: $\lim_{\mu \rightarrow 0} \xi_2 = 0$. Hence, the eigenvector is bounded for μ sufficiently small.

The eigenvector corresponding to λ_2 is found similarly:

$$\frac{a_{11}}{\mu} \eta_1 + \frac{a_{12}}{\mu} \eta_2 = \lambda_2 \eta_1$$

$$a_{21} \eta_1 + a_{22} \eta_2 = \lambda_2 \eta_2$$

$$\frac{a_{11}}{\mu} \eta_1 + \frac{a_{12}}{\mu} \eta_2 = \frac{\eta_1 [\mu a_{22} + a_{11} - \sqrt{(\mu a_{22} + a_{11})^2 - 4\mu \det A}]}{2\mu}$$

$$\frac{\eta_1}{\eta_2} = \frac{-2a_{12}}{a_{11} - \mu a_{22} + \sqrt{(\mu a_{22} + a_{11})^2 - 4\mu \det A}}$$

Letting $\eta_2 = 1$ we get the eigenvector $\begin{pmatrix} \eta_1 \\ 1 \end{pmatrix}$, and we note that for μ sufficiently small this vector is bounded.

Let

$$(6.9) \quad P = \begin{pmatrix} 1 & \eta_1 \\ \xi_2 & 1 \end{pmatrix};$$

then

$$(6.10) \quad P^{-1} = \frac{1}{1 - \eta_1 \xi_2} \begin{pmatrix} 1 & -\eta_1 \\ -\xi_2 & 1 \end{pmatrix}.$$

Clearly, for μ sufficiently small P^{-1} is bounded.

Also,

$$(6.11) \quad P^{-1} A_\mu P = D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

for μ sufficiently small.

The difference equation associated with system (6.4) is

$$\sum_{i=0}^k \alpha_i \begin{pmatrix} Y_{n+i} \\ Z_{n+i} \end{pmatrix} = h \sum_{i=0}^k \beta_i A_h \begin{pmatrix} Y_{n+i} \\ Z_{n+i} \end{pmatrix}.$$

Let

$$\begin{pmatrix} Y_{n+i} \\ Z_{n+i} \end{pmatrix} = P \begin{pmatrix} U_{n+i} \\ V_{n+i} \end{pmatrix}.$$

Substituting this expression into the difference equation and multiplying on the left by P^{-1} , we obtain

$$\begin{aligned} \sum_{i=0}^k \alpha_i \begin{pmatrix} U_{n+i} \\ V_{n+i} \end{pmatrix} &= h \sum_{i=0}^k \beta_i P^{-1} A_h P \begin{pmatrix} U_{n+i} \\ V_{n+i} \end{pmatrix} \\ &= h \sum_{i=0}^k \beta_i \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} U_{n+i} \\ V_{n+i} \end{pmatrix} \\ &= h \sum_{i=0}^k \beta_i \begin{pmatrix} \lambda_1 U_{n+i} \\ \lambda_2 V_{n+i} \end{pmatrix}. \end{aligned}$$

Recalling that

$$\lambda_1 \sim \frac{a_{11}}{h^\gamma} \quad (\text{as } h \rightarrow 0)$$

and

$$\lambda_2 \sim \frac{\det A}{a_{11}} \quad (\text{as } h \rightarrow 0)$$

and $a_{11} < 0$, we can conclude that μ -stability is a sufficient condition to ensure $|U_n| \rightarrow 0$ and $|V_n|$ remains

bounded as $h \rightarrow 0$, $x_n = nh$. Since P is a bounded matrix for h sufficiently small, $|Y_n|$ and $|Z_n|$ will remain bounded.

THEOREM 6.1. Let the consistent finite difference equation

$$(6.13) \quad \sum_{i=0}^k \alpha_i \begin{vmatrix} Y_{n+i} \\ Z_{n+i} \end{vmatrix} = h \sum_{i=0}^k \beta_i A_n \begin{vmatrix} Y_{n+i} \\ Z_{n+i} \end{vmatrix} + \begin{vmatrix} R_n^1 \\ R_n^2 \end{vmatrix}$$

where

$$A_n = \begin{vmatrix} \frac{a_{11}}{h^\gamma} & \frac{a_{12}}{h^\gamma} \\ a_{21} & a_{22} \end{vmatrix}$$

a constant matrix, satisfy the following conditions:

a) $R_n^1 = \mathcal{O}(h^{2(1-\gamma)})$ and $R_n^2 = \mathcal{O}(h^{2-\gamma})$ as $h \rightarrow 0$.

b) The finite difference equation is μ -stable.

c) $a_{11} < 0$

d) $Y_i - y(x_i) = \mathcal{O}(h^{1-\gamma})$ and $Z_i - z(x_i) = \mathcal{O}(h^{1-\gamma})$

as $h \rightarrow 0$ $i = 0, 1, \dots, k-1$.

Then the difference equation is convergent.

Proof: The difference equation satisfying the error terms $e_n = Y_n - y(x_n)$ and $d_n = Z_n - z(x_n)$ is

$$(6.14) \quad \sum_{i=0}^k \alpha_i \begin{vmatrix} e_{n+i} \\ d_{n+i} \end{vmatrix} = h \sum_{i=0}^k \beta_i A_h \begin{vmatrix} e_{n+i} \\ d_{n+i} \end{vmatrix} + \begin{vmatrix} R_n^1 - T_n^1 \\ R_n^2 - T_n^2 \end{vmatrix}$$

where the last term is $\begin{vmatrix} \mathcal{O}(h^{2(1-\gamma)}) \\ \mathcal{O}(h^{2-\gamma}) \end{vmatrix}$ as $h \rightarrow 0$.

Introducing

$$\begin{vmatrix} e_n \\ d_n \end{vmatrix} = P \begin{vmatrix} u_n \\ v_n \end{vmatrix}$$

where P was defined earlier and

$$P^{-1} A_h P = \begin{vmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{vmatrix}$$

where

$$\lambda_1 \sim \frac{a_{11}}{h} \quad \text{as } h \rightarrow 0$$

and

$$\lambda_2 \sim \frac{\det A}{a_{11}} \quad \text{as } h \rightarrow 0$$

we can write (6.14) as

$$(6.15) \quad (\alpha_i 1 - h\beta_i \begin{vmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{vmatrix}) \begin{vmatrix} u_{n+i} \\ v_{n+i} \end{vmatrix} = P^{-1} \begin{vmatrix} R_n^1 - T_n^1 \\ R_n^2 - T_n^2 \end{vmatrix} =$$

$$= \begin{vmatrix} \mathcal{O}(h^{2(1-\gamma)}) \\ \mathcal{O}(h^{2-\gamma}) \end{vmatrix}$$

where we have used the explicit form of $P^{-1} = \begin{vmatrix} \mathcal{O}(1) & \mathcal{O}(1) \\ \mathcal{O}(h^\gamma) & \mathcal{O}(1) \end{vmatrix}$

The equation for u_n can be treated like the one in Theorem 2.2, while the equation for v_n follows the standard ODE theory. Q.E.D.

THEOREM 6.2. Let the consistent finite difference equation

$$(6.16) \quad \sum_{i=0}^k \alpha_i \begin{vmatrix} Y_{n+i} \\ Z_{n+i} \end{vmatrix} = h \sum_{i=0}^{k'} \beta_i \begin{vmatrix} h^{-\gamma} F_{n+i} \\ G_{n+i} \end{vmatrix} + \begin{vmatrix} R_n^1 \\ R_n^2 \end{vmatrix}$$

where $F_n = f(Y_n, Z_n, x_n)$ and $G_n = g(Y_n, Z_n, x_n)$ and R_n^1 and R_n^2 are the round-off errors satisfy the following conditions:

a) $R_n^1 = \mathcal{O}(h^{2-2\gamma})$ and $R_n^2 = \mathcal{O}(h^{2-\gamma})$ as $h \rightarrow 0$.

b) The finite difference equation is relatively stable.

c) The first and second partial derivatives of f and g are continuous and bounded for $0 \leq x \leq 1$, $-\infty < y < \infty$, and $-\infty < z < \infty$. Furthermore, $-\bar{L} \leq \frac{\partial f}{\partial y} \leq -\bar{L} < 0$ in this strip.

d) $e_i = \mathcal{O}(h^{1-\gamma})$ and $d_i = \mathcal{O}(h^{1-\gamma})$ as $h \rightarrow 0$,
 $i = 0, 1, \dots, k-1$.

Then there exists a constant C for any h such that

$$|e_n| \leq C h^{1-\gamma} \quad \text{and} \quad |d_n| \leq C h^{1-\gamma} \quad \text{for } n = 0, 1, \dots, N$$

where $x_N = 1$.

Proof: The exact solution of the system (6.2) satisfies

$$(6.17) \quad \sum \alpha_i \begin{vmatrix} y_{n+i} \\ z_{n+i} \end{vmatrix} = h \sum_{i=0}^k \beta_i \begin{vmatrix} h^{-\gamma} f_{n+i} \\ g_{n+i} \end{vmatrix} + \begin{vmatrix} T_n^1 \\ T_n^2 \end{vmatrix}$$

where $f_{n+i} = f(y_{n+i}, z_{n+i}, x_{n+i})$ and $g_{n+i} = g(y_{n+i}, z_{n+i}, x_{n+i})$. $T_n^1 = \mathcal{O}(h^{2-2\gamma})$ and $T_n^2 = \mathcal{O}(h^{2-\gamma})$ as $h \rightarrow 0$ by consistency and condition c).

Subtracting (6.17) from (6.16) and letting $e_n = Y_n - y_n$ and $d_n = Z_n - z_n$ gives

$$(6.18) \quad \sum_{i=0}^k \alpha_i \begin{vmatrix} e_{n+i} \\ d_{n+i} \end{vmatrix} = h \sum_{i=0}^k \beta_i \begin{vmatrix} h^{-\gamma} (F_{n+i} - f_{n+i}) \\ G_{n+i} - g_{n+i} \end{vmatrix} + \begin{vmatrix} R_n^1 - T_n^1 \\ R_n^2 - T_n^2 \end{vmatrix}$$

$$= h \sum_{i=0}^k \beta_i \begin{vmatrix} h^{-\gamma} \bar{f}_{yn+i} & h^{-\gamma} \bar{f}_{zn+i} \\ \bar{g}_{yn+i} & \bar{g}_{zn+i} \end{vmatrix} \begin{vmatrix} e_{n+i} \\ d_{n+i} \end{vmatrix} + \begin{vmatrix} R_n^1 - T_n^1 \\ R_n^2 - T_n^2 \end{vmatrix}$$

Letting $N_0 = \frac{\ell n h^{1-\gamma}}{\ell n r}$ we can obtain a bound for the error terms in the standard way (see Theorems 4.1 and 5.2) for $0 \leq n \leq N_0$. Assuming this has been done now add

$$-h^{1-\gamma} \sum_{i=0}^k \beta_i \frac{\partial f(y(x_0), z(x_0), x_0)}{\partial y} e_{n+i} = h^{1-\gamma} \sum_{i=0}^k \beta_i L_0 e_{n+i}$$

to both sides of (6.18).

Note: If $\frac{\partial f}{\partial y}$ is a monotonic increasing function as is usually the case, then $L_0 = \bar{\bar{L}}$.

$$(6.19) \quad \left\| \sum_{i=0}^k (\alpha_i + h^{1-\gamma} \beta_i L_0) e_{n+i} \right\| = \left\| \sum_{i=0}^k \alpha_i d_{n+i} \right\|$$

$$\begin{aligned} & h \sum_{i=0}^k \beta_i \left\| \frac{h^{-\gamma} (L_0 + \bar{f}_{yn+i})}{\bar{g}_{yn+i}} \frac{h^{-\gamma} \bar{f}_{zn+i}}{\bar{g}_{zn+i}} \right\| \times \left\| \frac{e_{n+i}}{d_{n+i}} \right\| + \left\| \frac{R_n^1 - T_n^1}{R_n^2 - T_n^2} \right\| \\ & \equiv \left\| \begin{matrix} q_n^1 \\ q_n^2 \end{matrix} \right\| \end{aligned}$$

The solutions to (6.19) are

$$(6.20) \quad e_n = \begin{cases} \sum_{i=0}^{n-k} \bar{\Phi}_{n-i-1}^1 q_i^1 + \Psi_n^1 & n \geq k \\ \Psi_n^1 & n < k \end{cases}$$

where $\bar{\Phi}_n^1$ and Ψ_n^1 are defined like $\bar{\Phi}_n$ and Ψ_n in Theorem 4.1 and

$$(6.21) \quad d_n = \begin{cases} \sum_{i=0}^{n-k} \Phi_{n-i-1}^2 q_i^2 + \psi_n^2 & n \geq k \\ \psi_n^2 & n < k \end{cases}$$

where Φ_n^2 and ψ_n^2 are defined like g_n and θ_n in Theorem 2.2. By continuity and condition c) for some $\varepsilon > 0$

$$|L_0 + \bar{F}_{yn+i}| \leq \varepsilon + \mathcal{O}(E_{n+i} + D_{n+i})$$

for $0 \leq n \leq N_1 - k$ where $E_n = \max_{i=0,1,\dots,n} |e_i|$ and $D_n = \max_{i=0,1,\dots,n} |d_i|$.

Using (6.20) and following the procedure of Theorem 4.1 we can obtain a bound for $|e_n|$.

$$|e_n| \leq \varepsilon K_1 E_n + K_2 D_n + \mathcal{O}((E_n + D_n)E_n) + \mathcal{O}(h^{1-\gamma})$$

$$\text{for } N_0 \leq n \leq N_1$$

or if $\varepsilon K_1 < 1$, then

$$(6.22) \quad E_n \leq K_3 D_n + \mathcal{O}((E_n + D_n)E_n) + \mathcal{O}(h^{1-\gamma})$$

$$\text{for } 0 \leq n \leq N_1$$

where K_1 , K_2 , and K_3 are positive constants.

Letting $\max |\Phi_n^2| = \Phi^2$, we have from (6.21) that

$$\begin{aligned}
 |d_n| &\leq \Phi^2 h \sum_{i=0}^{n-k} \sum_{j=0}^k |\beta_j| (G_1 E_n + G_2 D_n) \\
 &\quad + \Phi^2 \sum_{i=0}^{n-k} \mathcal{O}(h^{2-\gamma}) + |\Psi_n^2| \\
 &\leq x \Phi^2 \sum_{j=0}^k |\beta_j| [G_1 E_n + G_2 D_n] + \mathcal{O}(h^{1-\gamma}) \\
 &\qquad\qquad\qquad \text{for } 0 \leq n \leq N_1
 \end{aligned}$$

where $|\bar{g}_{yn}| \leq G_1$, $|\bar{g}_{zn}| \leq G_2$, and $x_n = nh \leq x$.

Substituting (6.22) into (6.23) gives

$$|d_n| \leq x K_4 D_n + \mathcal{O}((E_n + D_n)E_n) + \mathcal{O}(h^{1-\gamma}).$$

Choosing x small enough so that $1 > x K_4$ we can write

$$(6.24) \quad D_n \leq \mathcal{O}(h^{1-\gamma}) + \mathcal{O}((E_n + D_n)E_n) \quad 0 \leq n \leq N_1$$

and consequently from (6.22)

$$(6.25) \quad E_n \leq \mathcal{O}(h^{1-\gamma}) + \mathcal{O}((E_n + D_n)E_n).$$

Adding (6.24) and (6.25) and letting $W_n = E_n + D_n$

we obtain

$$W_n \leq O(h^{1-\gamma}) + O(W_n^2).$$

The proof is completed like Theorem 4.1.

VII. APPLICATIONS AND FINAL STATEMENTS

A series of non-linear boundary layer problems of the kind discussed in Chapter 4 were solved on the IBM 7094 and the CDC 6600 computers by various finite difference schemes. At first γ was chosen equal to $\frac{1}{2}$ and mesh-sizes of $h = .01, .001, .0005, .0003, .0001$, and $.00005$. Although various different schemes were used to solve the same boundary layer equation, we concentrated on finite difference methods that illustrated the need for the new kind of stability called μ -stability, i.e., most of our schemes had two or more roots of $\rho(\zeta)$ located on the unit circle.

We also considered the cases of double roots of $\rho(\zeta)$ and other methods that were stable but not μ -stable. The latter methods produced divergent results, but the divergence was not clearly obvious until x_n was outside the boundary layer. Inside the boundary layer we observed two or three place accuracy with the exact solution or with the comparison with optimal methods that were μ -stable schemes.

Finally, we allowed γ to vary between 0 and 1 and observed the effect on convergent of μ -stable consistent schemes. We tried $\gamma = .1, .25, .5, .75$, and $.9$. We noted that for $.5 < \gamma < 1$ the iterative methods required much

more time and a few cases even required choosing smaller values of h . However, consistent μ -stable methods were observed to converge even for these values of γ .

The accuracy of the results varied with the precision of the difference methods. We found, for example, that Euler's method ($Y_{n+1} - Y_n = h^{1-\gamma} F(Y_n, x_n)$) gave only three to four place accuracy while optimal methods gave six or greater place accuracy when compared with known solutions.

In every case when h became sufficiently small schemes which were predicted to converge by the theory did so, while schemes which were predicted to diverge overflowed in the computer.

We illustrate some of the numerical results in the following graphs and tables. Graph 1 shows the numerical calculations in the boundary layer for a particular non-linear differential equation with three different choices for μ .

Graph 2 compares the numerical data from a μ -stable method with a stable method. In this case the μ -stability condition is that $\beta_1 < 1$. Notice that divergence doesn't occur with the stable (but not μ -stable) method inside the boundary layer, but outside of it.

Figures 1, 2, and 3 give the number of exact significant figures in the numerical data when compared to the exact answer calculated to eight significant figures. The

methods refer to the following numerical schemes:

Method 1 (Euler's Method)

$$Y_{n+1} - Y_n = h^{1/2} F(Y_n, x_n)$$

Method 2 (Adam's Method)

$$Y_{n+2} - Y_{n+1} = \frac{h^{1/2}}{12} (5F(Y_{n+2}, x_{n+2}) + 8F(Y_{n+1}, x_{n+1}) - F(Y_n, x_n))$$

Method 3 (μ -stable but not relatively stable)

$$Y_{n+2} - Y_n = h^{1/2} \left(\frac{3}{4} F_{n+2} + \frac{1}{2} F_{n+1} + \frac{3}{4} F_n \right)$$

Method 4 (all β_i 's ≥ 0 ; roots $\zeta = 1, -.2$)

$$Y_{n+2} - .8 Y_{n+1} - .2 Y_n = h^{1/2} (.4 F_{n+2} + .8 F_{n+1})$$

Method 5 (A poor μ -stable method)

$$Y_{n+2} - Y_n = h^{1/2} (.5 F_{n+2} + F_{n+1} + .5 F_n)$$

Method 6 (Simpson's Rule; divergent)

$$Y_{n+2} - Y_n = h^{1/2} \left(\frac{1}{3} F_{n+2} + \frac{4}{3} F_{n+1} + \frac{1}{3} F_n \right)$$

This data was calculated on the IBM 7094, and it was not as good as corresponding data calculated on the CDC 6600.

We suspect that round-off errors in the IBM machine caused the poorer showing.

Figure 4 compares Method 1 (Euler's Method) and Method 7--the optimal method for the roots $\zeta = 1, .5, .5$:

$$Y_{n+3} - 2 Y_{n+2} + 1.25 Y_{n+1} - .25 Y_n =$$

$$\frac{h^{1/2}}{96} (41 F_{n+3} + 19 F_{n+2} - 53 F_{n+1} + 17 F_n) .$$

This data was calculated on the CDC 6600.

Finally, in Figure 5 we compare the "worst methods" (large $|C_0|$ or large $\sum_{i=0}^k |\beta_i|$) with the "best methods" (those guaranteed to converge by Corollary 1 or Corollary 2 of Theorem 4.1).

Method 8 (large $\sum_{i=0}^k |\beta_i|$)

$$Y_{n+3} - Y_{n+2} = h^{1/2} (F_{n+3} + 10 F_{n+2} + 2 F_{n+1} - 12 F_n)$$

Method 9 (large $|C_0|$; roots $1, .9, .9$)

$$Y_{n+3} - 2.8 Y_{n+2} + 2.61 Y_{n+1} - .81 Y_n =$$

$$h^{1/2} (F_{n+3} + 3.01 F_{n+2} - F_{n+1} - 3 F_n)$$

Methods 10 to 19

The optimal methods for the roots $\zeta = 1, -.i, -.i$ and $i = 0, 1, \dots, 9$.

For $i = 0$ and 1 Corollary 2 can be used to guarantee convergence.

For $i = 2, 3, \dots, 9$ all β_i 's ≥ 0 . Hence, by Corollary 1 convergence is assured.

CHART OF SIGNIFICANT FIGURES

$$f(y,x) = -y(y - 1)(20x + 10); \quad y(0) = 2$$

$$h = .001 \quad \gamma = .5$$

x-coordinate	Method						Comment
	1	2	3	4	5	6	
.001	3	6	4	6	6	6	
.005	3	6	5	6	5	7	
.010	4	5	5	5	5	7	
.015	4	4	4	4	7	7	
.020	4	6	6	6	6	7	
.025	5	5	6	5	6	7	
.035	6	6	6	6	6	8	
.045	7	7	7	7	5	4	
.055	8	8	8	7	5	4	End of Boundary Layer
.065	8	8	8	7	5	3	
.100	8	8	8	7	5	2	
.200	8	8	8	7	5	0	Divergence in 6
.300	8	8	8	7			Instability in 5
.400	8	8	8	7			
.500	8	8	8	7			
.600	8	8	8	7			
.700	8	8	8	7			
.800	8	8	8	7			
.900	8	8	8	7			
1.00	8	8	8	7			

Figure 1 Data from IBM 7094

CHART OF SIGNIFICANT FIGURES

$$f(y,x) = -y(y - 1)(20x + 10); \quad y(0) = 2$$

$$h = .0001 \quad \gamma = .5$$

x-coordinate	Method				Comment
	1	2	3	4	
.0001	4	6	6	6	
.0005	3	6	6	6	
.0010	4	6	5	6	
.0015	4	6	6	5	
.0020	4	5	5	5	
.0025	5	6	6	5	
.0030	5	6	6	5	
.0035	4	6	6	5	
.0040	5	5	6	5	
.0045	5	7	6	5	
.0050	5	5	5	5	
.0055	5	5	6	5	
.0060	5	5	6	5	
.0065	4	4	6	4	
.0075	5	5	6	5	
.0095	5	5	6	5	
.0150	5	7	7	6	End Boundary Layer
.0250	8	8	8	7	
.1	8	8	8	7	
.2	8	8	8	7	
.3	8	8	8	7	
.4	8	8	8	7	

Figure 2 Data from IBM 7094

CHART OF SIGNIFICANT FIGURES

$$f(y,x) = -y(y - 1)(20x + 10); \quad y(0) = 2$$

$$h = .00005 \quad \gamma = .5$$

x-coordinate	Method				Comment
	1	2	3	4	
.00005	4	6	6	6	
.00055	4	6	6	4	
.001	4	5	5	5	
.0015	5	5	6	5	
.0025	5	6	6	5	
.0035	5	5	5	5	
.0045	5	5	5	5	
.0060	5	5	6	5	
.01	7	7	7	6	
.15	8	8	8	7	End Boundary Layer
.20	8	8	8	7	
.25	8	8	8	7	
.30	8	8	8	7	
.35	8	8	8	7	
.40	8	8	8	7	
.45	8	8	8	7	
.50	8	8	8	7	
.55	8	8	8	7	
.60	8	8	8	7	
.65	8	8	8	7	
.70	8	8	8	7	
.75	8	8	8	7	

Figure 3

Data from IBM 709⁴

CHART OF SIGNIFICANT FIGURES

$$f(y,x) = -y(y - 1)(20x + 10); \quad y(0) = 2$$

$$h = .0001 \quad \gamma = .5$$

x-coordinate	Method		Comment
	1	7	
.0001	4	9	
.0005	4	9	
.001	4	9	
.002	3	9	
.005	5	9	
.01	7	9	End Boundary Layer
.02	9	9	
.05	9	9	
.1	9	9	
.15	9	9	
.20	9	9	
.25	9	9	
.30	9	9	
.35	9	9	
.40	9	9	
.45	9	9	
.50	9	9	
.55	9	9	
.60	9	9	

Figure 4

Data from CDC 6600

CHART OF SIGNIFICANT FIGURES

$$f(y,x) = -y(y - 1)(20x + 10); \quad y(0) = 2$$

x-coordinate	Method			Comment
	8	9	10-19	
.0001	1	1	9	
.0005	2	1	9	
.001	2	1	9	
.005	4	3	9	
.01	6	3	9	
.015	2	3	9	End Boundary Layer
.02	3	3	9	
.03	0	3	9	Divergence in 8
.1	0	1	9	
.2		0	9	Divergence in 9
.3		0	9	
.4			9	
.5			9	
.6			9	
.7			9	
.8			9	
.9			9	
1.0			9	

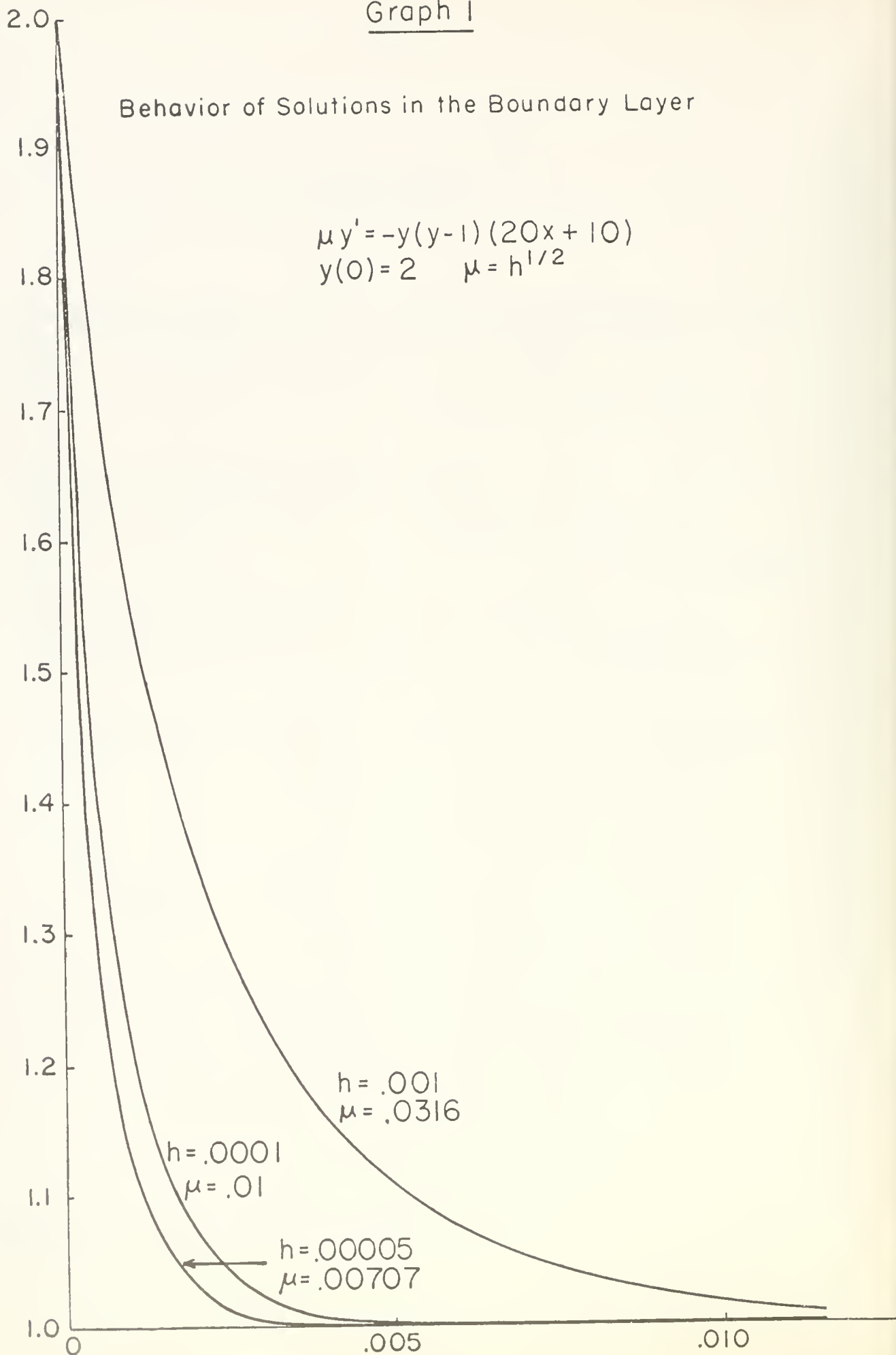
Figure 5 Data from CDC 6600

BIBLIOGRAPHY

- [1] Dahlquist (1956): Convergence and stability in the numerical integration of ordinary differential equations. Math. Scan., 4, 33-53.
- [2] Henrici (1962): Discrete Variable Methods in Ordinary Differential Equations, John Wiley and Sons, New York.
- [3] Hull and Luxemburg (1960): Numerical methods and existence theorems for ordinary differential equations. Num. Math., 2, 30-41.
- [4] Vasil'eva (1963): Asymptotic behavior of solutions. Russian Math. Surveys, 18, 13-84.
- [5] Hull and Newbery (1959): Error bounds for a family of three-point integration procedures. J. Soc. Indust. Appl. Math., 7, 402-412.
- [6] Hull and Newbery (1961): Procedures which minimize propagated errors. J. Soc. Indust. Appl. Math., 9, 28-47.

Graph 1

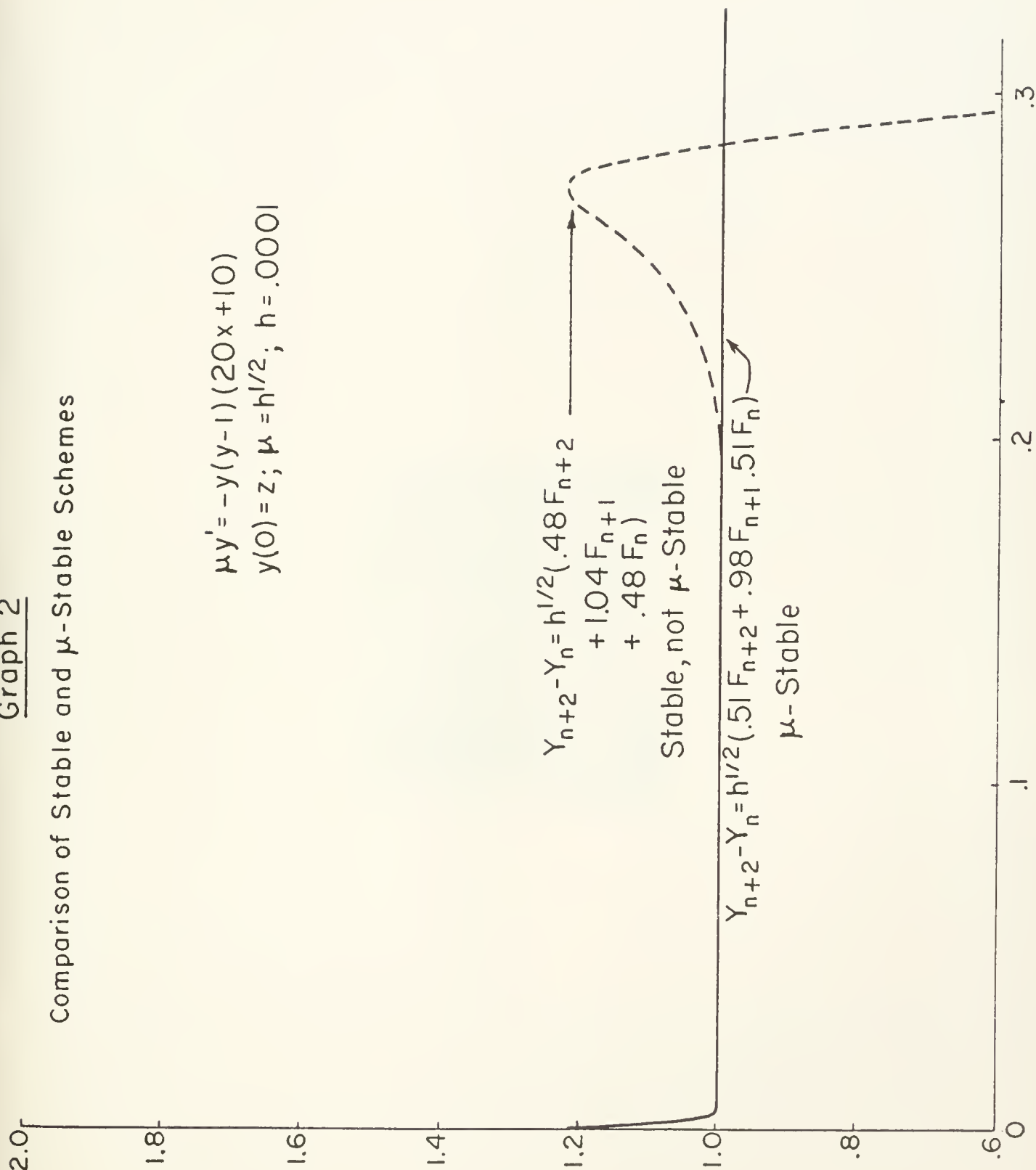
Behavior of Solutions in the Boundary Layer



Graph 2

Comparison of Stable and μ -Stable Schemes

$$\begin{aligned}\mu y' &= -y(y-1)(20x+10) \\ y(0) &= z; \mu = h^{1/2}; h = .0001\end{aligned}$$



This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

- A. Makes any warranty or representation, express or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or
- B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission, or employee of such contractor, to the extent that such employee or contractor of the Commission, or employee of such contractor prepares, disseminates, or provides access to, any information pursuant to his employment or contract with the Commission, or his employment with such contractor.

SEP 21 1966
DATE DUE

APR 17 '67

GAYLORD

PRINTED IN U.S.A.

NYU

NYO-

1480-63

Murphy

c.1

NYU

NYO-

1480-63 Murphy

c.1

NYU

NYO-

1480-63 Murphy

c.1

AUTHOR

Numerical analysis of

TITLE

boundary layer problems.

DATE DUE

BORROWER'S NAME

APR 17 '67

Oliver Kline

N.Y.U. Courant Institute of
Mathematical Sciences
251 Mercer St.
New York 12, N. Y.

